Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 336 (2005) 175–189
BBRC
www.elsevier.com/locate/ybbrc

The repertoire of solute carriers of family 6: Identification of new human and rodent genes

Pär J. Höglund, Dijana Adzic, Sara J. Scicluna, Jonas Lindblom, Robert Fredriksson *

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

Received 3 August 2005
Available online 18 August 2005

Abstract

Tremendous amount of primary sequence information has been made available from the genome sequencing projects, although a complete annotation and identification of all genes is still far from being complete. Here, we present the identification of two new human genes from the pharmacologically important family of transporter proteins, solute carriers family 6 (SLC6). These were named SLC6A17 and SLC6A18 by HUGO. The human repertoire of SLC6 proteins now consists of 19 functional members and four pseudogenes. We also identified the corresponding orthologues and additional genes from mouse and rat genomes. Detailed phylogenetic analysis of the entire family of SLC6 proteins in mammals shows that this family can be divided into four subgroups. We used Hidden Markov Models for these subgroups and identified in total 430 unique SLC6 proteins from 10 animal, one plant, two fungi, and 196 bacterial genomes. It is evident that SLC6 proteins are present in both animals and bacteria, and that three of the four subfamilies of mammalian SLC6 proteins are present in *Caenorhabditis elegans*, showing that these subfamilies are evolutionary very ancient. Moreover, we performed tissue localization studies on the entire family of SLC6 proteins on a panel of 15 rat tissues and further, the expression of three of the new genes was studied using quantitative real-time PCR showing expression in multiple central and peripheral tissues. This paper presents an overall overview of the gene repertoire of the SLC6 gene family and its expression profile in rats.

Keywords: Genome projects; Profile hidden Markov models; SLC6; Evolution

Genes coding for membrane proteins are one of the largest groups of genes in vertebrate genomes. Membrane proteins are claimed to constitute more than 10% of all genes in the human [1] and mouse genomes [2], respectively. Important subgroups of membrane proteins are G-protein coupled receptors [3,4], single pass tyrosine kinase receptors [5,6], ion channels [7], and transporters or solute carriers. Solute carriers control the uptake and flow of various substances such as sugar, amino acids, nucleotides, inorganic ions, and drugs over the cell membrane, either passively or actively [8]. There are currently 43 different families of solute carriers [9] and one of these is the solute carrier family 6 (SLC6). To separate the letter 6 of the solute carrier type from the family member number a capital A is introduced, i.e., SLC6A4 (solute carrier family 6) separates the gene family name "6" from the family member number "4." The SLC6 family of proteins acts as specific transporters for neurotransmitters, amino acids, and osmolytes like betaine, taurine, and creatine. The neurotransmitters whose levels are controlled by SLC6 proteins are low molecular weight molecules like acetylcholine, dopamine, serotonin, norepinephrine, and amino acids like γ-aminobutyric acid (GABA), glycine, and aspartate. The SLC6 proteins are Na^+ cotransporters and the energy for the transport of the solute against its concentration gradient is provided by the electrochemical gradient for sodium ions. In some cases, such

* Corresponding author. Fax: +46 18 51 15 40.
E-mail address: Robert.fredrikson@neuro.uu.se (R. Fredriksson).

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.08.048

as for the SLC6A1 (GAT1), chloride ions are also cotransported, although this is variable between the members of the family. In addition, the serotonin transporter SLC6A4 requires the counter transport of potassium ions but the role of chloride or potassium ions in this transport process is not well understood.

SLC6 family members are also known to be important for a number of pathological conditions and several of them are potential drug targets that are pursued by the pharmaceutical industry. Mutations in the SLC6A2 (norepinephrine transporter) cause orthostatic intolerance. Orthostatic intolerance is a syndrome characterized by lightheadedness, fatigue, and syncope, and is associated with postural tachycardia and plasma norepinephrine concentrations that are disproportionately high in relation to sympathetic outflow. The dopamine transporter SLC6A3 (DAT1) that acts to take released dopamine back up into presynaptic terminals has been implicated in human disorders such as parkinsonism, Tourette syndrome, and substance abuse. The serotonin transporter encoded by SLC6A4 is the target of an important class of antidepressant drugs, the serotonin selective reuptake inhibitors. It has been shown to take part in numerous clinical conditions such as autism, depression, neuroticism, and obsessive compulsive disorder. There is a high prevalence of SLC6A8 mutations in X-linked mental retardation. Common features of X-linked mental retardation are neurological disturbances including seizures, behavioral problems, speech delay, and inability to engage in structured play [10]. Mutations in the predominantly neutral amino acid transporter SLC6A19 are the cause of Hartnup disorder. The SLC6A14 gene encodes an amino acid transporter, which potentially regulates tryptophan availability for serotonin synthesis and thus possibly affects appetite control and mood. Recently, a Finnish–Swedish study has shown that polymorphism in the SLC6A14 gene is associated with obesity. A replication study with obese French subjects reconfirmed this conclusion [11,12].

Cloning of the human transporters and their rodent orthologues; SLC6A1 (GABA), (SLC6A2) (norepinephrine/dopamine), SLC6A3 (dopamine), and SLC6A4 (serotonin) in 1990–1994 marked the beginning of the identification of genes in the SLC6 family [13–16]. Additional genes that code for monoamine and amino acid carriers were identified during the late 1990s. Still, several additional genes belonging to this family are ''orphans'' or without known functions [17,18]. The common structural feature of SLC6 transporters is the presence of a putative 12 transmembrane (TM) region with the N- and C-terminal regions located inside the plasma membrane. The transporters are all relatively large and complex proteins, and consist of between 590 and 800 amino acid residues. SLC6 proteins have not yet been crystallized and hence the precise structure of the SLC6 neurotransmitter is not very well determined.

There has been confusion regarding the nomenclature, naming of orthologous proteins in different species, and the repertoire of functional proteins for the SLC6A family of proteins. For example, the GABA neurotransmitter transporter 2 (SLC6A13) in human is orthologous to the GABA 3 transporter in mouse. The human creatine transporter type 2 (SLC6A10) has been localized to 16p11.2 by two studies [19,20]. This paper and one other study have identified the SLC6A10 gene as a pseudogene with an early stop codon [21]. However, this gene has been considered to be a functional gene both in Unigene at NCBI as well as in a recent review article [8]. It is known that the family SLC6 is ancient and exists in eukaryotes as well as in eubacteria and archaebacteria [22]. However, evolutionary studies of the SLC6 family are scarce. In 1998, it was concluded that this gene family existed in insects, bacteria, and *Caenorhabditis elegans* but not in yeast and fungi [22]. Tremendous amount of primary sequence information has become available from recent sequencing projects, providing a near to full coverage of the entire genomes from a diversity of animal, plant, and bacteria species, although a complete annotation and identification of all gene is still far from complete [23] and new genes are continuously being described [24–26]. The complete genome material makes comprehensive analysis of the gene repertoire in multiple genomes possible.

In this study, we aim to provide a complete view of the gene repertoire of the SLC6 family of proteins. We accomplished this through searches in the human, mouse, and rat genomes using various bioinformatic methods. Several new members of the SLC6 family were found and we present genomic structures, orthologous relationships, phylogenetic grouping, and EST expression charts for the entire SLC6 family. Moreover, we conducted an evolutionary study on the SLC6 repertoire in 13 different species. We also performed reverse transcriptase PCR on a tissue panel of rat mRNA to provide a quantitative expression profile of the entire SLC6 family.

Results

Our initial strategy was to download all known members of the human, mouse, and rat SLC6 family. Known human SLC6 sequences were identified on the HUGO Gene Nomenclature Committee Homepage as well as from GenBank (http://www.ncbi.nih.gov/Genbank/). Known mouse sequences were identified on the Mouse Genome Informatics web page and from GenBank. The rat genome database RatMap contained only six SLC6 sequences and therefore we also searched Rat Genome Database (http://rgd.mcw.edu/). The protein sequences were thereafter downloaded from the Entrez web page (http://www.ncbi.nlm.nih.gov/entrez/). The

known transporters retrieved were the human SLC6A1-SLCA16 and SLC6A20, the mouse transporters mSLC6A1-A9, mSLC6A10-15 as well as mSLC6A20, and the rat transporters rSLC6A1-A9, rSLC6A10-15, and rSLC6A18. The previously unknown gene SLC6A19 was released and added to our data set during the finalization of the study [27,28].

To investigate whether additional SLC6 genes exist in human and rodents, a profile HMM was downloaded from http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00209. This profile was used to identify SLC6 like sequences using HMMsearch from the 2.3.2 release of the HMMER package [29]. We also used BLAST-searches with known SLC6 proteins as baits. The searches were performed in various databases as specified under Materials and methods. This resulted in the identification of two new human genes. We approached the HUGO Genome Nomenclature committee and these were confirmed to be novel and were named SLC6A17 and SLC6A18 to adhere to the accepted nomenclature. SLC6A18 was present in the sequenced part (21,243 cDNA clones) of a library in a large-scale cDNA sequencing project and automatically annotated as Xtrp2 [30], but was not characterized further. We also found one mouse gene and three rat genes, and these were named mSLC6A19, and rSLC6A14, rSLC6A19, and rSLC6A20 according to the orthologous human SLC6 family member. The repertoire in human also contains two previously reported pseudogenes and during our investigation one additional pseudogene, SLC6A10pB, was identified. The gene repertoire in human, mouse, and rat consists of 19 genes in each species and that mouse and rat displays an identical 1:1 phylogenetic relationship. On the contrary, the repertoire in rodents and human is not identical as SLC6A16 does not exist in mouse and rat while the X transporter protein 3 similar 1 (Xtrp3s1) does exist not in human (see Fig. 1).

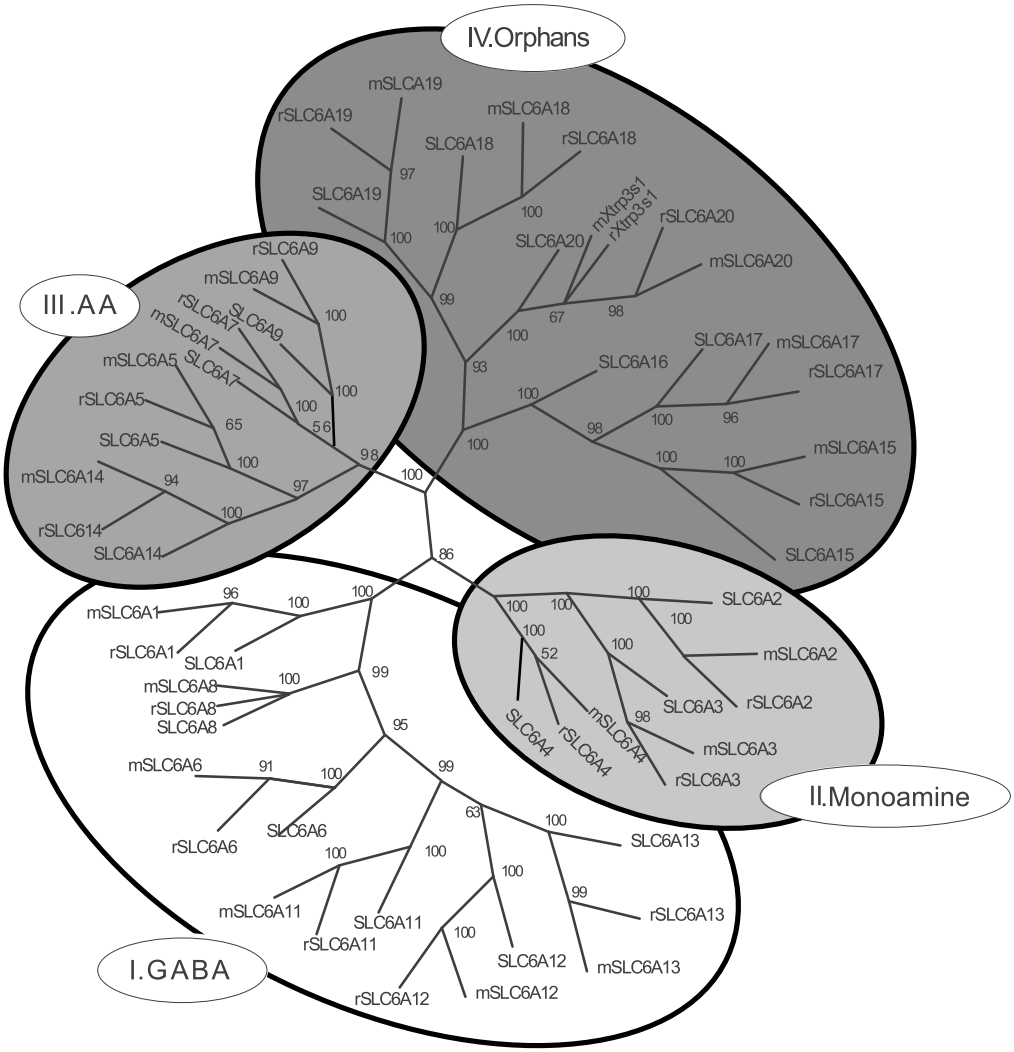

Fig. 1. A phylogenetic analysis of the solute carriers family 6 (SLC6) in human, mouse and rats. Maximum likelihood with 100 replicas was used to calculate phylogenetic trees. The family groups into five subgroups: I, GABA; II, Monamine; III, Amino Acid; and IV, Orphans.

The new genes identified from the HMM-searches were found as GENSCAN gene predictions. GENSCAN and other ab initio gene prediction programs are known to detect around 80% of all exons, although the exact exon–intron boundaries are almost newer correctly predicted. Therefore, before phylogenetic analysis, the new genes have to be manually assembled using other nongenomic information. Below we show the origin and how each protein was assembled.

Human

SLC6A17 (DJ1003J2_NTT4) was found through HMM-searches. It had a total length of 436 amino acids, which is shorter than other members of the family. We extended the sequence by using the previously known 724 amino acid long mouse ortholog (NP_758475) as bait in BLAT searches against the human genome. We extracted the corresponding human DNA sequence and used it to search for human and mouse ESTs and cDNA. We used 18 human EST sequences to correct splice site predictions. There were still gaps in the protein and therefore we added five mouse cDNA and 11 mouse ESTs, which resulted in the entire sequence of human SLC6A17 being verified. The human SLC6A17 has a 97% and 97.6% identity to mouse and rat, respectively.

SLC6A18 was found through searches with BLAST using SLC6A16 as bait. The identified sequence was gi|34785074 (FLJ31236). In addition, 1 human cDNA and 7 human EST sequences were assembled, which covered the whole sequence. The SLC6A18 has 82.4% and 82.5% amino acid identity to the mouse and rat SLC6A18, respectively.

SLC6A21p (Hs19_11347_28_95_2) was found when searching with the HMM model against the GENSCAN assembly 28. It had a length of 420 amino acids and is located at 19q13.33 next to SLC6A16. The length of known SLC6 proteins ranges from 591 (rSLC6A20) to 799 amino acids (rSLC6A5). The consensus sequence for the PFAM00209 (sodium: neurotransmitter symporter family) model has a length of 536 amino acids. Searches in all species for ESTs and mRNA yielded no hits. The short length and the absence of any mRNA or EST make it likely that SLC6A21 is a pseudogene.

SLC6A10p is a recent duplicate of the creatinin transporter SLC6A8 protein that was identified by Iyer et al. [19]. It is located at 16p11.2 and it is concluded that there was duplication between 16p11.1 and Xq28. SLC6A8 is located at Xq28. The duplication probably occurred within recent evolutionary time (7–10 mya) [21]. Predicted translations of exons and RT-PCR analyses revealed that this paralog to SLC6A10 on chromosome 16 probably is a pseudogene as there is a one base substitution in exon 4 from Trp (TGG) to a stop codon (TGA) [21]. While searching the May 2004 assembly of the human genome using BLAT at the UCSC website, we found that there are two adjacent pseudogenes about 890 kb apart, SLC6A10pA (32797531–32799840) and SLC6A10pB (33690486–33692794). We downloaded these genome sequences from UCSC, produced alignment between these genomic pieces with ClustalW, and used Megalign to calculate a nucleotide percentage identity. We saw that these two pseudogenes have a 99.6% identity. SLC6A8 has a 96.8 and 97.0 percentage nucleotide identity to SLC6A10p (32797531–32799840) and SLC6A10p (33690486–33692794), respectively.

Mouse

Mouse SLC6A16 (XP_145587) was derived from a GNOMON gene prediction, which was further supported by EST evidence. Our searches yielded 11 ESTs and 1 cDNA sequence, and all of these were expressed in testis. 94 out of 728 amino acids, including three splice sites were not covered by EST or cDNA. BLAT searches revealed that there is a local duplication of SLC6A16 in both mouse and rat. The genes are located at 7qB2, only 5 kb apart. These recently duplicated genes in both mouse and rat are most likely pseudogenes and were therefore excluded from our analysis.

mSLC6A19 has one cDNA and two EST sequences that cover the whole protein. During the writing process, another group cloned this protein and linked it to the Hartnup disease [31].

Rat

The rSLC6A14 (XP_233305) was inspected using alignments with the human and mouse orthologues of SLC6A14. Six rat and 18 mouse ESTs aligned. The sequence and the exon–intron-boundaries were verified to be correctly predicted.

The rSLC6A19 was derived using a GNOMON gene prediction. In the original prediction, it was 718 amino acids with 14 exons. It was corrected using three rat EST, two mouse cDNA, and 38 mouse EST sequences. The final sequence is a 634 amino acid protein with 12 exons. The exon–intron boundaries were corrected by the cDNA, mRNA, and mouse, human alignment.

The rat SLC6A16 was found by searching the predicted GNOMON rat dataset based on the NCBI build 2. There is a local duplication of the SLC6A16 pseudogene. The pseudogenes are located on 1q22 and are 9 kb apart. The use of RT-PCR revealed that the rSLC6A16 is not expressed in our rat tissue panel. The lack of expression supports the conclusion that the SLC6A16 duplicate is a pseudogene.

The rSLC6A20 (XP_217302.2) was corrected by alignment with the human and mouse orthologue as well as rat ESTs. Searching with BLAT shows that the gene is located adjacent to rXtrp3s (Q64093). The relation-

ship is the same in mouse where rXtrp3s is located next to mSLC6A20.

Phylogeny

Phylogenetic analysis with maximum likelihood, neighbor-joining and maximum parsimony tree construction methods clearly divides the SLC6 family into four subgroups (Fig. 1). One hundred bootstrap replicas were used in each method. We have chosen to designate these subgroups GABA (bootstrap value: 100, 100, and 52), Monoamine (100, 100, and 94), Amino Acid (98, 97, and 94) and Orphans (98, 100, and 94) according to substrate preferences. These high bootstrap values and the fact that all methods suggest the same topology support this subdivision of the SLC6 family into four clades.

EST expression pattern in mouse and human

The human and mouse EST databases were downloaded and BLAST was performed with the SLC6 proteins as baits. The human and mouse EST sequences were collected and then searched with BLASTX (translated query versus protein database) back against a human and mouse RefSeq dataset obtained from http://www.ncbi.nlm.nih.gov/. Rat was excluded, as very few ESTs are available from this species. RefSeq is an annotated collection of protein sequences and we had added our novel gene products to this protein dataset to ensure that each EST would hit the best corresponding protein. This procedure creates a detailed and specific expression profile for the SLC6 family (Table 2). Five transporters have a high expression (10 or more ESTs) in the CNS (mSLC6A1, mSLC6A7, SLC6A8, mSLC6A8, and mSLC6A9) and five transporters have a high expression level in the eye (mSLC6A1, mSLC6A6, SLC6A8, and mSLC6A9, mSLC6A15). Also the *Monoamine* transporters SLC6A2, A3, and A4 are, as expected, mainly expressed in the CNS, although their expression levels seem to be relatively low (1, 6, and 2 ESTs for the mouse proteins). The SLC6A2 (16 ESTs) is the only transporter with a high expression in female reproductive organs and this seems to be specific for human, as no ESTs are detected from this tissue in mouse. mSLC6A6 is the only transporter that is highly expressed in the male reproductive organ (20 ESTs) while zero ESTs are detected for this transcript from testis in human. Three transporters (mSLC6A8, mSLC6A13, and mSLC6A19) have a high expression in the gastrointestinal system and the five (mSLC6A8, mSLC6A13, mSLC6A18, mSLC6A19, and mSLC6A20) are highly expressed in the kidney. These transporters belong to the GABA (mSLC6A8 and mSLC6A13) or *Orphan* (mSLC6A18, mSLC6A19, and mSLC6A20) family. It should be noted that the GABA family member SLC6A8 has a broad expression pattern with high expression in melanocytes and in the CNS. The SLC6A8 is also very highly expressed in tumor tissues (102 ESTs). Although there are numerous cases in Table 2 where the expression pattern is similar between human and mouse, there are a surprisingly high number of cases where the expression pattern differs remarkably. Some of these differences can probably be attributed to differences in tissue sampling between the two species, but many cases are not likely to be explained by this. One such example is the absence of the expression of SLC6A1 and SLC6A9 in human eye, while several other transporters such as SLCA8 and SLCA15 have a similar expression in the eye of the two species. This indicates that there is an actual difference in the expression levels that is unlikely to be triggered due to sampling differences.

To obtain an expression profile over the entire family of SLC6 proteins in rat, we used reverse-transcriptase (RT) PCR on mRNA from 14 different rat tissues (Table 3). The PCR products were analyzed using agarose gel electrophoresis and each sample was analyzed twice where both replicas had to show the same result to be classified as positive. To avoid genomic contamination, all RNA was treated with DNAse and tested for genomic contamination by excluding RT from the reactions. All cDNAs were tested and verified to work with GAPDH primers and all cDNAs in the panel were diluted to show a similar amount of GAPDH expression. Three of these were CNS tissues, mainly brainstem, cerebellum, and forebrain. Clear expression in the brain was seen for 7 of the 20 transporter proteins but only three *Orphan* transporters (SLC6A15, SLCA18, and SLCA19) were found to be abundant in forebrain. Interestingly, the proteins showing high expression in the forebrain are the two phylogenetically closely related *Orphan* transporters SLCA18 and SLCA19. SLC6A6 is the only one of the proteins that is expressed in female reproductive organs and our EST analysis shows that this is also true for male reproductive organs in mouse.

The expression pattern of three phylogenetically related orphan transporters SLC6A18, SLC6A19, and SLC6A20, two of which were shown to be expressed in brain from our RT-PCR panel, was further tested for expression in 13 defined sub-dissected regions of rat brain as well as a number of peripheral tissues using real-time quantitative RT-PCR (qPCR). This method has a higher sensitivity and requires smaller amounts of material than ordinary RT-PCR, making it more suitable for sub-dissected brain regions. The qPCR method also allows for a quantitative analysis while ordinary RT-PCR is at best semi-quantitative. In Fig. 2, we display the amount of the given transcript relative the amount of GAPDH in 13 brain-regions and eight peripheral tissues. SLC6A18 and SLC6A19 are expressed in dorsal, but not ventral, hypothalamus at relatively high levels (50% of GAPDH) while SLC6A20, as

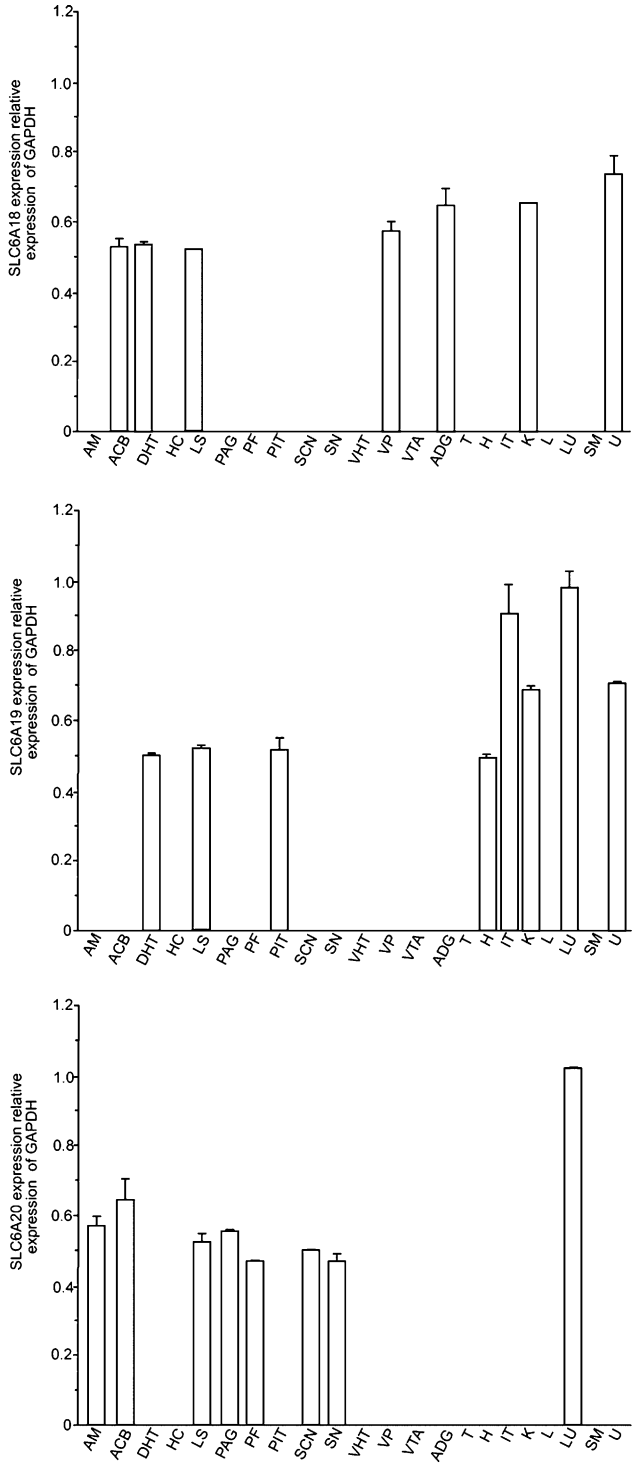

Fig. 2. Expression level of SLC6A18, SLC6A19, and SLC6A20 in different rat tissues relative to GAPDH. Data are presented as means ± SEM ($n = 3$) as a fraction of GAPDH expression. *Abbreviations:* AM, amydala; ACB, acumbens; DHT, dorsal hypothalamus; HC, hippocampus; LS, locus ceruleus; PAG, periaqueductal gray area; PF, prefrontal cortex; PIT, pituitary; SCN, suprachiasmatic nucleus; SN substantia nigra; VHT, ventral hypothalamus; VP, ventral pallidum; VTA, ventral tegmental area; ADG, adrenal gland; T, testis; H, heart; IT, intestins; K, kidney; L, liver; LU, lung; SM, skeletal muscle; U, uterus.

expected from our RT-PCR analysis, does not seem to be expressed in either of the hypothalamic regions. On the other hand, certain defined brain regions do have a substantial expression of this protein, something we were unable to detect in our larger brain regions. This is probably due to dilution effects in the less defined regions.

Evolutionary analysis

We downloaded 13 predicted protein datasets: *Takifugu rubripes*, *Danio rerio*, *Ciona intestinalis*, *C. elegans*, *Caenorhabditis briggsae*, *Drosophila melanogaster*, *Anopheles gambiae*, *Arabidopsis thaliana*, *Schizosaccharomyces pombe*, and *Saccharomyces cerevisiae*. They were searched with the SLC6 HMM and the hits were retrieved. These proteins were then searched by BLAST towards an adjusted human Refseq including our novel human transporters. The proteins were divided into the GABA, Monoamine, Amino Acid, and Orphan subgroups according to the BLAST hits, and all proteins that did not place into any of these groups were considered unclassifiable. To be classified as SLC6 proteins, the five best hits for each protein had to be a SLC6 protein and four hits had to belong to a specific subgroup in order to be classified as a member of that subgroup. The result of the evolutionary analysis is given in Fig. 3. The numbers of proteins in different groups are similar in mouse, rat, human, and chicken. The *T. rubripes*, *D. rerio*, and *C. intestinalis* all had a relatively high number of GABA transporters. The *A. gambiae* and *D. melanogaster* both had the highest number of transporters in the *Amino Acid* group whereas most of the *C. elegans* proteins are *Unclassifiable*. The plant (*A. thaliana*) and yeast (*S. cerevisiae* and *S. pombe*) predicted protein datasets do not seem to contain any SLC6-protein. The predicted protein datasets from the 196 available bacterial genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) yielded 16 GABA, 4 Monoamine, 3 Amino Acid, 3 Orphans, and 81 Unclassifiable proteins.

Discussion

Our thorough mining of the entire genome of human, mouse, and rat using HMM led to identification of two new human SLC6 proteins. We also identified six previously unidentified rodent proteins. We used these new proteins together with the previously known proteins from this family and performed the first overall evolutionary analysis for the entire SLC6-family in mouse, rat, and human. We also investigated the number of SLC6 sequences in protein prediction datasets from an additional 13 eukaryotic species. In addition to the phylogenetic and sequence analysis, we also performed an expression analysis over the entire family of SLC6

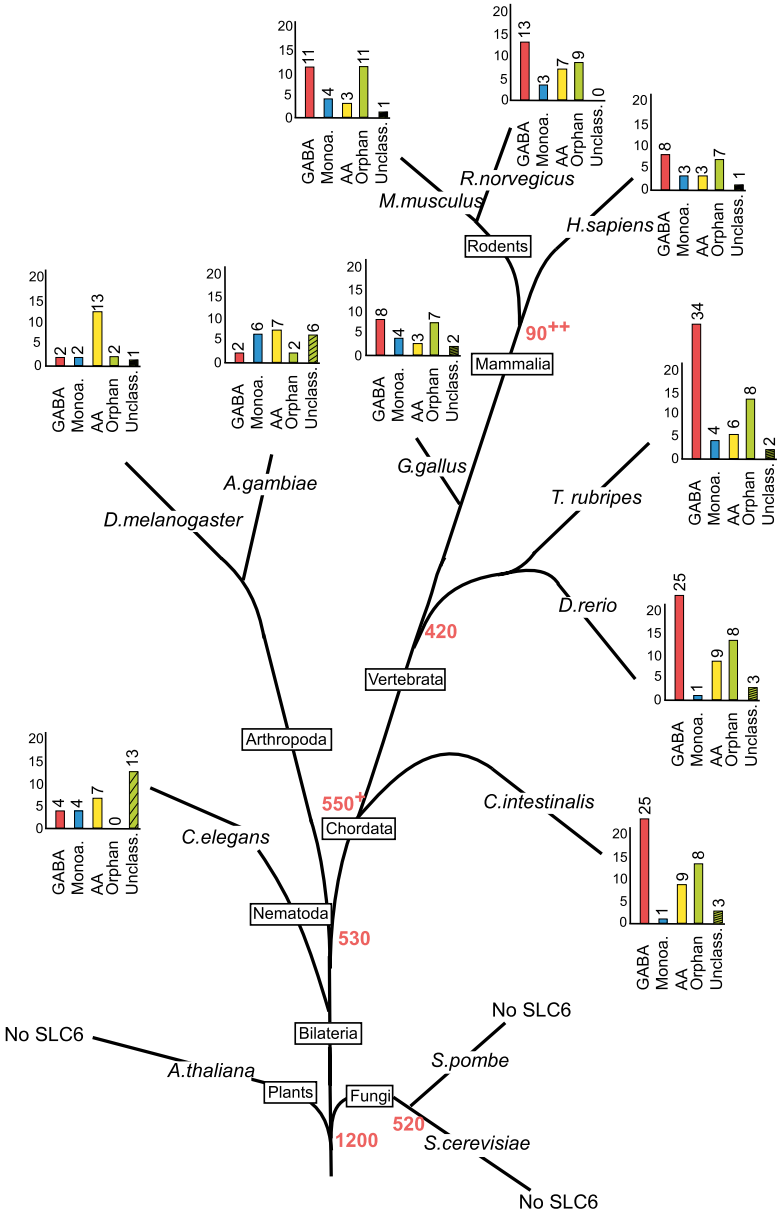

Fig. 3. Evolutionary tree with the number of solute carriers family 6 (SLC6) in different species indicated in graphs. The graph displays the number of SLC6 genes at the *Y*-axis and the main class above the *X*-axis. The number at the nodes indicates the time in million years since the split at that node occurred, based on fossil data according to [38] [+][39] [++][40]. Phylogenetic analyses and subsequent BLAST divides the gene products into five subgroups: GABA, Monoamine (Monoa.), Amino Acid (AA), Orphans, and Unclassifiable.

proteins, in human and mouse based on EST data and in rat by RT-PCR and qPCR.

Tremendous amount of primary sequence information from the human, mouse, and rat genomes has been released, providing an almost full coverage of the respective genomes. We can therefore conclude that the gene repertoires presented in Table 1 are fairly complete and that this is likely to be a complete set of genes from the SLC6 family in human, mouse, and rat. The completion of the sequencing projects has shifted the focus from sequencing and initial protein prediction to the discovery and annotation of new genes and understanding of their function, structure, and evolution. Our analysis shows that the SLC6 family in mouse and rat is identical. Human and rodents have similar repertoire, although there are differences in the *Orphan* group. SLC6A16 exists in the human, but not in rodents. The X transporter protein 3 similar 1(Xtrp3s1) is a duplication of SLC6A20 that only exists in mouse and rat. Here we present two new genes, SLC6A17 and SLC6A18, that belong to the *Orphan* subgroup. SLC6A18 forms a phylogenetic cluster with SLC6A19 and SLCA20 (Fig. 1), while SLC6A17 groups with SLC6A15 and SLC6A16. Our expression studies, based on EST, RT-

Table 1
Description of the complete repertoire of the SLC6 family in human, mouse, and rat

Name	Alias	Length (aa)	Accession No.	Chromosomal position [UCSC]	Exons
SLC6A1	GAT, GABATR, GABATHG	599	NP_003033	3p25.3	14
SLC6A2	NET, NAT1, NET1	617	NP_001034	16q12.2	13
SLC6A3	DAT, DAT1	620	NP_001035	5p15.33	13
SLC6A4	5-HTT, SERT	630	NP_001036	17q11.2	13
SLC6A5	NET1, GLYT2	797	NP_004202	11p15.1	15
SLC6A6	TAUT	619	NP_003034	3p25.1	13
SLC6A6p				21q21.1	
SLC6A7	PROT	636	NP_055043	5q33.1	13
SLC6A8	CT1; CRTR	635	NP_005620	Xq28	13
SLC6A9	GLYT1	692	NP_964012	1p34.1	14
SLC6A10pA				16p11.2	
SLC6A10pB				16p11.2	
SLC6A11	GAT3; GAT-3	632	NP_055044	3p25.3	14
SLC6A12	BGT1; BGT-1	614	NP_003035	12p13.33	14
SLC6A13	GAT2; GAT-2	602	NP_057699	12p13.33	14
SLC6A14	OBX; ATB(0+)	642	NP_009162	Xq23	14
SLC6A15	V7-3; NTT73; hv7-3	730	NP_877499	12q21.31	11
SLC6A16	NTT5	737	NP_054756	19q13.33	12
SLC6A17	NTT4	727	XP_371280*	1p13.3	11
SLC6A18	Xtrp2; FLJ31236	628	NP_872438	5p15.33	12
SLC6A19		634	NP_001003841	5p15.33	12
SLC6A20	XT3; Xtrp3	592	NP_071800	3p21.31	11
SLC6A21p				19q13.33	
mSLC6A1	Gat1; XT-1; Gabt1; Xtrp1	599	P31648	6qE3	14
mSLC6A2	NET	617	NP_033235	8qC5	13
mSLC6A3	DAT, DAT1	619	NP_034150	13qC1	12
mSLC6A4	Htt; Sert; 5-HTT	630	NP_034614	11qB5	13
mSLC6A5	Glyt2	791	NP_683733	7qB3	15
mSLC6A6	Taut	621	NP_033346	6qD1	13
mSLC6A7	Prot	637	NP_958741	18qE1	13
mSLC6A8	CRT; CT1; CRTR; Creat	640	NP_598748	XqA7.1	13
mSLC6A9	Glyt1; Glyt-1	633	NP_032161	4qD2.1	12
mSLC6A11	GAT4	627	B44409	6QE3	14
mSLC6A12	GAT2; Gabt2	614	NP_598422	6qF1	14
mSLC6A13	GAT3; Gabt3	602	NP_653095	6qF1	14
mSLC6A14	ATB0plus; CATB0plus	638	NP_064433	XqA2	14
mSLC6A15		729	NP_780537	10qD1	11
mSLC6A17	NTT4	724	NP_758475	3QF2.3	11
mSLC6A18	XT2; Xtrp2	615	NP_035860	13qC1	12
mSLC6A19	B<0>AT1	634	XP_127449	13qC1	12
mSLC6A20		635	NP_035861	9qF4	11
mXtrp3s1		592	NP_631881	9qF4	11
rSLC6A1	RGD:620533, Gabt1	599	NP_077347	4q42	14
rSLC6A2	RGD:621822, Net	597	NP_112633.1	19p11	14
rSLC6A3	RGD:3715, Dat1	619	NP_036826	1p11	13
rSLC6A4	RGD:3714	653	NP_037166	10q26	15
rSLC6A5	RGD:621824, GLYT2	799	NP_976079	1q22	14
rSLC6A6	RGD:61912	621	NP_058902	4q34	13
rSLC6A7	RGD:620928, Prot	661	NP_446448	18q12.1	13
rSLC6A8	RGD:619711, CHOT1	635	NP_059044	Xq37	13
rSLC6A9	RGD:621243, Glyt1	633	NP_446270	5q36	12
rSLC6A11	RGD:628737, Gabt4, Gat3	627	NP_077348	4q42	14
rSLC6A12	RGD:620255, RNU28927	628	NP_059031	4q42	11
rSLC6A13	RGD:620788, GAT-2	602	NP_598307	4a42	14
rSLC6A14		640	XP_233305	Xq12	14
rSLC6A15	RGD:628664, Ntt73	729	NP_758824	7q21	11
rSLC6A17		727	P31662	*	*
rSLC6A18	RGD:69352, Rosit, Xtrp2	615	NP_058859	1p11	12
rSLC6A19		634		1p11	12
rSLC6A20		591		8q32	11
Xtrp3s1		616	Q64093	8q32	11

The columns are name, alias, length in amino acids, accession number, and chromosomal position (from the UCSC website) and the total number of exons. The 'r' prefix denotes rat and the 'm' prefix indicates mouse. The p suffix as in for instance SLC6A10pA shows that it is a pseudogene. A * means that the information is not available due to missing genomic information.

Table 2
Description of the expressed sequence tags (EST) expression in mouse (A) and human (B)

	CNS	Eye	Female reproduction	Gastrointestine	Kidney	Male reproduction	Melanocyte	Other	Stem cells	Tumor
A										
mSLC6A1	44(30)	12(1)						1		
mSLC6A2	1							8(8)		
mSLC6A3	6(1)							1		
mSLC6A4	2(2)		4	1				2(1)		
mSLC6A5	3									
mSLC6A6	7(3)	13	2	1	10	20(8)			1	8
mSLC6A7	17(2)								1	
mSLC6A8	10(5)	14(10)	7	11	10			19(12)	3	12
mSLC6A9	12(2)	14(2)		4(1)				7(4)		5
mSLC6A11	5(1)	1		1				2(1)		
mSLC6A12	1		4	8	1			2(2)		
mSLC6A13	7(2)	2		19	15	2(2)		4(2)		
mSLC6A14	1(1)		1	8				5(4)		1
mSLC6A15	9(4)	10(7)						4(3)		3
mSLC6A17	8(6)	2		1(1)	1			1		
mSLC6A18				3	60			1		
mSLCA19	1			11	16			6(1)		
mSLC6A20				1	41	2		3(1)		
mXtrp3s1	7	1		2	4			1		
B										
SLC6A1	9			1(1)					1(1)	1(1)
SLC6A2			16					3		12
SLC6A3			1							1
SLC6A4			7							
SLC6A5	1							3		
SLC6A6	1	1	3					13	3	3
SLC6A7	3									
SLC6A8	13(1)	14	8	5	2	7	13	30(2)	6	102
SLC6A9			2	2				4	1	14
SLC6A11	4(1)						1	25		1
SLC6A12	2		3	1		1				4
SLC6A13	3(1)	8		4	5			12	1	1
SLC6A14			1	5				12		7
SLC6A15	2(2)	4	3				2	7	2	6
SLC6A16	7					7		8(1)	1	4
SLC6A17	1		2							4
SLC6A18	1			2	4					
SLC6A19										

The parenthesis shows the number ESTs from that category found in fetal tissue.

PCR, and qPCR, show that transcripts from the two novel genes, as well as their phylogenetic neighbors, are present in a number of both central and peripheral tissues, indicating that these genes are likely to participate in several important physiological functions.

The phylogenetic analysis (Fig. 1) divides the SLC6 family into four subgroups which fit remarkably well with the known substrate preferences for the SLC6 proteins, and we have hence opted for naming the phylogenetic branches accordingly. Our phylogenetic analysis reveals that the human SLC6 repertoire consists of 6 GABA, 3 Monoamine, 4 Amino Acids, and 6 Orphan proteins, in total of 19 SLC6 proteins. The bootstrap values that designate the four groups are generally high, regardless of the phylogenetic method which shows that the phylogenetic groupings are very stable. The values are between 97 and 100 for the four nodes in all cases, with the exception of the *GABA* subgroup in the maximum parsimony analysis where it is 52. It is a well-known fact that maximum parsimony generally performs worse than the other methods for larger datasets when the degree of sequence identity is low. Interestingly, the *Amino Acid* substrate subgroup is phylogenetically closer to the *Orphan* subgroup and recently the SLC6A19 gene from the *Orphan* subgroup has been found to transport neutral amino acids [27,28]. Hence, considering that the phylogenetic clustering for all other groups is according to substrate preference, it is possible that other *Orphan* SLC6 transporters could have amino acids as substrates. As some of these *Orphan* transporters are highly expressed in brain, it is tempting to speculate that they function as transporters for amino acids into the cells that are subsequently used as substrates for synthesis of neurotransmitters.

Table 3
RT-PCR analysis of the known SLC6 family on a panel of mRNA from 15 rat tissues

	Cerebellum	Brainstem	Forebrain	Uterus	Ovary	Adrenal gland	Fat tissue	Kidney	Lung	Heart	Liver	Testis	Skeletal muscle	Spleen	Intestine
rSLC6A1	■									+		+			
rSLC6A2				+				■	■				+	+	+
rSLC6A3										+		+	■		
rSLC6A4					+	+	+		■	■	+	■	■	+	+
rSLC6A5		■				+									
rSLC6A6	■	■	+	■	■	+	+		■	■	■	+	■	■	■
rSLC6A7															
rSLC6A8	+	+				+		■	■		■	+	■	■	■
rSLC6A9	■	■	+			■				+	+		+		
rSLC6A11													■		
rSLC6A12					+										
rSLC6A13		+			+	+									
rSLC6A14							+				+				
rSLC6A15	■	■	■	+	+	+	+		■	+	+				
rSLC6A17	+	+	+		+		+			+					
rSLC6A18	■	■	■	+		+	+								
rSLC6A19	■	■	■	+		■			+	■	+	■	+		■
rSLC6A20							+		+						
rXtrpP3s1	■		+		+							■			

Dark shaded areas indicate strong expression, whereas + indicates lower expression levels as determined by visual inspection on agarose gels.

Our evolutionary analysis based on GENSCAN protein predictions (Fig. 3) resulted in 22 SLC6 proteins in human; 8 GABA, 3 Monoamine, 3 Amino Acids, and 7 Orphans, and 1 Unclassified. The total number of known human genes is actually 22, but it is divided as; 6 GABA, 3 Monoamine, 3 Amino Acids, and 6 Orphans. We investigated this discrepancy with the phylogenetic analysis and chromosomal alignments. In 12 cases, there is a 1:1 match between the GENSCAN predictions and known genes, in three cases two GENSCAN genes partially match one known gene, i.e., GENSCAN has erroneously divided one gene into two predictions. In addition, 4 GENSCAN genes matched known pseudogenes and four of the known genes were not present in the gene prediction dataset. There is on average about 14 introns per human SLC6 protein, and this complex exon–intron arrangement makes it hard for ab initio programs such as GENSCAN to make correct predictions. This kind of analysis of GENSCAN genes, especially in gene families with a high number of introns, can be used to estimate the proportions of proteins in each subgroup but not to predict the absolute number of genes.

Our analysis shows that SLC6 genes are present in all animal species investigated, but not in plants or fungi. The repertoire of SLC6A proteins, as estimated from GENSCAN proteins, for mouse, human, rat, and chicken are very similar. However the *T. rubripes*, *D. rerio*, and *C. intestinalis* all had a 2- to 3-fold higher number of transporters in the GABA group. The *A. gambiae* and *D. melanogaster* both had most transporters in the *Amino Acid* group while *C. elegans* had the highest number in the unclassifiable group. The plant (*A. thaliana*) and yeast (*S. cerevisiae and S. pombe*) predicted protein dataset did not contain any SLC6-proteins while bacteria had a substantial number. A total of 16 GABA, 4 Monoamine, 3 Amino Acid, 3 Orphans, and 81 Unclassifiable bacterial proteins were found in the different genomes. The repertoire of SLC6 proteins in *C. elegans* and insects is rather different than in mammals. These do not have any genes belonging to the *Orphan* subgroup, which suggests that this subgroup has appeared after the divergence of the arthropoda line from the lineage leading to vertebrates. Although these lines do not have an Orphan subgroup the other three groups, GABA, Monoamine, and Amino Acids, are clearly present. Notably, there exist a few *Unclassifiable* genes in the genome of most species. For example there are 6 *Unclassifiable* genes in *A. gambiae* and 13 in *C. elegans* and 81 in bacteria. This is very intriguing and it is likely that one or more phylogenetic subgroups either appeared specifically in these lineages or existed in parallel with the three subgroups common to vertebrates and invertebrates, but were lost before the appearance of vertebrates.

In rat, nine of the 19 SLC6 proteins have high expression in brain, while three have low expression there. Of the known monoamine transporters for dopamine (SLC6A3), serotonin (SLC6A4), and noradrenaline (SLC6A2) only the serotonin transporter can be detected in large brain regions while all three are likely be detected in more defined regions. The three phylogenetically related GABA transporters SLC6A1, SLC6A8 and SLC6A6 were found in the CNS, while the other three proteins from the GABA group, SLC6A11, SLC6A12, and SLC6A13, seem to be less highly ex-

pressed and have more expression in the periphery. Expression in the periphery for rat GABA transporters has previously been shown for SLCA12 [32] and SLC6A13 [33]. In a similar manner, we do notice a high expression of all three monoamine transporters in peripheral tissues in rat, which is also supported by the EST data from human and mouse. This could be one reason for the side effects caused by antidepressant drugs in these tissues. The Amino Acid transporters SLC6A5 and SLC6A9 are expressed in brain and peripheral tissue, while SLC6A7 and SLC6A14 are expressed only in the periphery. Recently, polymorphisms in the SLC6A14 gene were shown to be involved in inherited forms of obesity [11,12].The gene is expressed in the gastrointestinal region in human and mouse with EST analysis and in fat tissue in rat. The transporters from the Orphan group are so called because the majority of them do not have any known substrate. Recently, human and mouse SLC6A19 was shown to be able to transport neutral amino acids [27] and mutations in this gene have been shown to be the cause of the renal disease Hartnup disorder. This disease is an autosomal recessive disorder which results from impaired transport of neutral amino acids across epithelial cells in renal proximal tubules and intestinal mucosa. It has also been proposed that the allelic heterogeneity of the *Orphan* transporter SLC6A16 may explain the variation in the biochemical and clinical phenotype in pedigrees with Hartnup disorder [28]. The SLC6A19 in rat shows expression in several peripheral tissues including intestine and in mouse this gene seems to have high expression in intestine and kidney. On the other hand, our human EST panels indicate that this gene has a very low expression level as not a single EST is found in any tissue. Also, we do not find any evidence of expression of the SLC6A16 protein in colon or kidney. The fact that the gene for SLC6A16 is absent in both mouse and rat complicates further the use of rodents as a model for this disease. The *Orphan* transporter SLC6A18 is highly expressed in brain and we can also detect SLC6A20 in certain subdissected brain regions of rat. Also SLC6A15 and SLC6A17 as well as the rodent specific Xtrp3s1, are expressed in the brain of all species investigated. This is intriguing as some of these proteins could have important functions in the brain such as acting as transporters for amino acids used in the synthesis of neurotransmitters or being specific reuptake proteins for neurotransmitters which currently do not have a known such system, for example trace amines.

In this analysis, we have provided a comprehensive overview of the repertoire of SLC6 genes in a number of species. This provides the molecular foundation for one of the important groups of targets for drugs that act on the transport uptake and flow of various neurotransmitters, amino acids, and osmolytes. We have identified new genes and provide convincing evidence that they code for functional proteins. Moreover, we provide expression profiles for the entire repertoire SLC6 mRNA in rats which may give better insight into the possible physiological role of these genes.

Materials and methods

Collecting the original dataset

Seventeen known human SLC6 sequences were initially identified on the HUGO Gene Nomenclature Committee Homepage (http://www.gene.ucl.ac.uk/nomenclature/). Fourteen mouse sequences were initially identified on the Mouse Genome Informatics web page (http://www.informatics.jax.org/). The rat genome database RatMap (http://ratmap.gen.gu.se) only contained six SLC6 sequences and therefore we also searched Rat Genome Database (http://rgd.mcw.edu/SSS) which contained eight more genes which resulted in a total of 14 unique rat genes. All the rat, mouse, and human protein sequences were thereafter downloaded from the Entrez web page (http://www.ncbi.nlm.nih.gov/entrez/). We also searched UniGene on the NCBI website and the literature for known genes.

Identification of novel genes

The collected SLC6 proteins were used as baits for searches in both the human, mouse, and rat protein databases. Information was collected from both NCBI (http://www.ncbi.nih.gov/) and the Celera Genomics (http://www.celeradiscoverysystem.com/) databases. The N- and C-termini were removed after being identified through searches against the RPS-BLAST database (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). The truncated genes were aligned using ClustalW 1.83 with default parameters applied [34]. From the alignment, a HMM model was constructed with hmmbuild from the 2.3.2 release of the HMMER package [29]. The model was calibrated using hmmcalibrate with default parameters. Searching was conducted in three human, one rat, and two mouse databases with hmmsearch using $E = 1e{-}4$ as the cut-off value. For human we searched the Gnomon protein data set (Build 34 version 3) and the GENSCAN protein dataset (assembly 28), both downloaded from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/). In addition, we also searched the GENSCAN dataset from Ensembl based on the NCBI version 34 assembly. For mouse, we used the Gnomon protein dataset of the NCBI Mouse build 32 as well as the ab initio Ensembl (http://www.ensembl.org/) version based on the NCBI Mouse build 32. For rat, we searched the Gnomon protein dataset, which is based on the NCBI build 2 from the Rat Genome Sequencing consortium (RGSC) v3.1 assembly.

We used protein–protein (BLASTP) searches of the non-redundant protein database at NCBI (http://www.ncbi.nih.gov/BLAST/), with the previously downloaded SLC6-genes as baits [35]. BLAST searches against human, mouse, and rat proteins were performed, without filtering for low complexity. The best five alignments for each search were manually inspected.

All the sequences were searched using BLAT against the human, mouse, and rat database at UCSC Genome Bioinformatics Site (http://genome.ucsc.edu). This was carried to both a species-to-species manner as well as across species. All high-scoring hits were visually inspected in order to find previously unidentified family members of the SLC6 family.

Verification of novel genes

The novel transporters were searched with BLASTP and RPS-BLAST. All genes with a SLC6 gene as the best BLASTP hit were subsequently searched with RPS-BLAST. The search with RPS-

BLAST had to yield a PFAM00209 (sodium:neurotransmitter symporter family) model as the best hit. All proteins that did not match both these two criteria were excluded.

Identification of EST-clones

The GENSCAN software, used for predicting coding regions for the human genome project, has the capacity to predict approximately 80% of the splice sites correctly [36]. Therefore, the coding regions needed to be verified. We used mRNA and expressed sequence tags (ESTs) from mouse, human, and rat, respectively. BLAT searches at the UCSC web site were used to obtain genomic nucleotide sequences for the novel protein sequences. The full genomic DNA sequences of the novel GPCRs were searched against the human and mouse NCBI EST databases, respectively. There is no rat-specific database and therefore we searched the entire EST database. The search was limited by the Entrez query "*Rattus norvegicus* [ORGN]". BLASTN against the nucleotide non-redundant database with the limit biomol_mrna[properties] was used to find the corresponding mRNA sequences. The alignments with the identified (ESTs) and cDNA sequences were manually inspected to ensure correct identity and exclude non-relevant hits. The predicted coding regions were verified by assembling the EST sequences and the full genomic DNA sequences using SeqMan 5.01 from the DNASTAR package. The genomic DNA sequences were considered correct, and the EST and mRNA sequences were only used to correct the predicted exon–intron boundaries. EditSeq 5.02 from DNASTAR package was used to find the right open reading frames and translate the nucleotide sequences into final novel protein sequences.

Phylogenetic analysis

Our protein sequences were all aligned with ClustalW version 1.83 [34]. The alignments were bootstrapped 100 times with SEQBOOT from the Linux version of the PHYLIP 3.6 package [37]. Three different methods, neighbor-joining, maximum parsimony, and maximum likelihood, were used to make phylogenetic trees from the SEQBOOT outfile. For the neighbor-joining method, protein distances were calculated with PROTDIST using the Jones–Taylor–Thornton matrix. The trees were calculated on the 100 different distance matrixes previously generated with PROTDIST, using NEIGHBOR. CONSENSE from the Linux version of the PHYLIP 3.6 package was used to obtain a bootstrapped consensus tree.

Maximum parsimony trees were calculated from the previously derived SEQBOOT file using PROTPARS from the Linux version of the PHYLIP 3.6 package. The trees were unrooted and calculated using ordinary parsimony and the topologies were obtained using the built-in tree search procedure. CONSENSE was used to obtain a bootstrapped consensus tree. Maximum likelihood trees were calculated, from the previously derived SEQBOOT file, using PROML from the Linux version of the PHYLIP 3.6 package. PROML was used with default settings and CONSENSE was used to obtain a bootstrapped consensus tree. All trees were plotted using TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

Expression profile

The EST database at NCBI currently has a total of 24.6 million entries. 6.0 million, 4.3 million, and 0.7 million are human, mouse, and rat EST sequences, respectively (http://www.ncbi.nlm.nih.gov/dbEST/index.html). The rat EST database was considered too small for making a rat EST expression profile. We searched the whole human and mouse EST databases with our SLC6 protein sequences as input files. We used protein query versus translated database, TBLASTN, from the BLAST 2.2.9 package. A cut-off value of 0.1 was used. The human and mouse EST sequences were collected and then searched with BLASTX (translated query versus protein database) and thereafter searched with BLASTX against a human and mouse RefSeq (ftp://ftp.ncbi.nih.gov/refseq/) datasets, respectively. We had added our novel transporters to these protein datasets to ensure that each EST would hit the best corresponding protein. We removed all ESTs that did not have a SLC6-gene as first hit. We then used in-house scripts to extract relevant expression information from the EST database. The ESTs were sorted and entered into a spreadsheet for manual inspection.

Description and retrieval of predicted protein datasets

The human, mouse, rat, *D. rerio*, *G. gallus*, *A. gambiae*, *T. rubripes*, *D. melanogaster*, and *C. elegans* predicted protein datasets were all retrieved from the Ensembl website. Ensembl predicted peptides are available in two different sets. One is the "pep" version, which consists of the super-set of all translations from known, ab initio or novel gene predictions. This set uses biological information such as EST-support to produce several different transcripts from each gene. The other set is the ab initio dataset with all predictions based entirely on the genomic sequences and no biological evidence. Prediction algorithms, such as SNAP and GENSCAN, have been used to create the datasets. The ab initio approach does not distinguish between pseudogenes and biologically real proteins. We however decided to use ab initio predicted protein datsets. Ab initio is a more honest and unbiased approach, because the result of the analyzes is not influenced by the amount of ESTs or any other biological data available. Thus, the ab initio set is better to use in a cross-species comparison. Ab initio also seems like a superior dataset in our test runs. The *D. melanogaster* dataset did only have a "pep" version. We used this version, but removed the multiple translated protein version and left only one transcript for each gene. The sources for the datasets for the yeast and bacterial genomes as well as the datasets for *A. thaliana* and *C. intestinalis* are indicated below. All other datasets were downloaded from the Ensemble site. Below follows a description of these datasets.

Anopheles gambiae. The ensemble *A. gambiae* (mosquito) genome release 17.2a.1 is a re-annotation of the second version of the *A. gambiae* genome assembly.

Arabidopsis thaliana. The *Arabidopsis* predicted protein dataset was downloaded from ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/.

Bacteria. The predicted protein datasets from the 196 available bacterial genomes were downloaded from ftp://ftp.ncbi.nih.gov/genomes/Bacteria/. The genomes have been searched with three prediction methods; GeneMark, GeneMark.hmm, and Glimmer. Fourteen genomes have also been reviewed by NCBI experts: *Aeropyrum pernix*, *Archaeoglobus fulgidus*, *Buchnera sp*, *Buchnera aphidicola*, *Corynebacterium glutamicum*, *Haemophilus influenzae*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Oceanobacillus iheyensis*, *Shewanella oneidensis*, *Pyrococcus abyssi*, *Pyrococcus horikoshii*, *Thermoplasma volcanium*, and *Vibrio vulnificus*.

Caenorhabditis elegans. The v25.116a.1 ensembl predicted protein dataset of *C. elegans* is a direct import of the Wormbase (http://www.wormbase.org/) 116 dataset. No additional gene building procedures have been preformed.

Gallus gallus. The chicken (*G. gallus*) retrieved is the ensembl 25.1b.1 version, based on the Chicken 1a build.

Ciona intestinales. The ciona proteins were downloaded from http://genome.jgi-psf.org.

Danio rerio. The genome of zebrafish (*D. rerio*) used was the ensembl zebrafish release 24.4.1 based on the zebrafish assembly version 4 (Zv4) produced by the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/D_rerio/).

Drosophila melanogaster. The ensembl 25.3b.1 version of *D. melanogaster* genome is based on the the Flybase (http://www.flybase.org/) 3.1 assembly. It is not available as an ab initio dataset.

Homo sapiens. We used the human ab initio ensembl v21.34d.1 build which is based on the NCBI 34 assembly of the human genome.

Mus musculus. We used the mouse ab initio release 24.33.1 based on the NCBI m33 mouse assembly.

Rattus norvegicus. We used the ensembl ab initio rat v25.3c.1 assembly. This version is based on the draft genome assembly (v3.1) found on the Baylor College of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu/projects/rat/).

Saccharomyces cerevisiae. The budding yeast (*S. cerevisiae*) has a current assembly available from the Sanger institute ftp-site (ftp://ftp.sanger.ac.uk/pub/yeast/cerevisiae/cerpep/). The predicted protein dataset consists of 5803 proteins.

Schizosaccharomyces pombe. The fission yeast (*S. pombe*) is also available from Sanger institute ftp-site(ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Protein_data/). The GENSCAN dataset consists of 4983 proteins.

Takifugu rubripes. The ensembl fugu (*T. rubripes*) genome release v25.2c.1 is based on the second Fugu assembly from the Fugu Consortium (http://www.fugu-sg.org/project/info.html).

Identification of SLC6 sequences in the evolutionary analysis

The predicted protein datasets previously retrieved were used. These predicted protein datasets were searched against a HMM-model (pfam00209.11) downloaded from http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00209 and searched using HMMsearch from the 2.3.2 release of the HMMER package using $E = 10$ as cut-off value. The hits in the result file were used to extract Fasta files, with an in-house script based on fastacmd from the NCBI BLAST-suite, from the original protein prediction file.

Subdivision of the SLC6 proteins

We divided our collection of mouse, human, and rat SLC6 proteins into five groups: (I) GABA & others (GABA), (II) Monoamine, (III) Amino acid, (IV) Orphan, and (V) Unclassifiable. This was done according to the phylogenetic analyses previously described. A file with all human RefSeq genes was downloaded from the ncbi ftp site: ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/. The RefSeq (Reference Sequence) protein is an annotated collection of non-redundant proteins. We thereafter added our confirmed 19 mouse and 19 rat proteins, and adjusted the RefSeq to make the human SLC6 repertoire look exactly as in our own dataset. We thereafter searched all our predicted proteins from the HMM search with BLASTP with an *e*-value of e−12. To be classified as SLC6 proteins, the five best hits for each protein had to be a SLC6 protein. The first four hits had to belong to a specific subgroup in order to be classified as a member of the subgroup. If, for instance, the first three BLAST hits were GABA proteins and hits four and five were *Monoamine*, the protein was concluded to be a *Unclassifiable* SLC6 protein.

RNA isolation and cDNA synthesis

Individual tissue samples were homogenized by sonication in TRIzol reagent (Invitrogen) using a Branson sonifier. Chloroform was added to the homogenate, which was then centrifuged at 12,000 rpm for 15 min at 4 °C. The water phase was transferred to a new tube, and RNA was precipitated with isopropanol. The pellets were cleaned with 75% ethanol, air-dried at room temperature, and dissolved in RNAse-free water. DNA contamination was removed by treatment with DNAse (Roche Diagnostics) for 12 h at 16 °C. The DNAse was then inactivated by heating the samples to 75 °C for 15 min. RNA purity was checked by PCR, using a primer for GAPDH (see below), and RNA concentration was determined with an Ultrospec 2100 spectrophotometer (Amersham Biosciences). cDNA was synthesized with reverse transcriptase from Amersham Biotech, using random hexamers as primers and following the instructions of the manufacturer. The quality of the cDNA was confirmed with PCR as above.

Reverse transcriptase PCR

The presence of individual transcripts in certain tissues was determined using reverse transcriptase (RT) PCR. RT-PCR was performed in a 25 μl final reaction volume using a Perkin Elmer 9700 thermal cycler (Applied Biosystems, Stockholm, Sweden). The PCR included 20 mM Tris–HCl (pH 8.4), 50 mM KCl, 4 mM $MgCl_2$, and 0.2 mM dNTP. Template concentration was 4 ng/μl, and the concentration of each primer was 1 pmol/μl. Taq DNA polymerase (Invitrogen) was used at 0.04 U/μl. Annealing temperature was 55 °C and 35 cycles were performed. All experiments were repeated twice. The PCR products were analyzed using agarose gel electrophoresis and scored manually. A certain tissue/primer pair combination had to be positive on both occasions to be considered positive. All PCR-primers were designed using the Primer 3 (http://frodo.wi.mit.edu/primer3/primer3_code.html) software using default settings, with a required product length between 80 and 110 nucleotides. All primer sequences are available from the authors upon request.

Real-time quantitative reverse transcriptase PCR

Relative levels of mRNA were determined by quantitative reverse transcriptase PCR (qPCR). qPCR was performed in a 25 μl final reaction volume using an iCycler real-time detection instrument (Bio-Rad Laboratories, Sundbyberg, Sweden). The reaction contained 20 mM Tris–HCl (pH 8.4), 50 mM KCl, 4 mM $MgCl_2$, 0.2 mM dNTP, and SYBR Green (1:50,000). Template concentration was 1 ng/μl, and the concentration of each primer was 0.8 pmol/μl. Taq DNA polymerase (Invitrogen) was used at 0.02 U/μl. All samples were measured in triplicate and compared with a no-template control for each primer pair. Annealing temperature was 62 °C and 50 cycles were performed. Melting point curves were included to confirm that only one product was formed. For a product to be classified as positive the CT-value for that product had to be two units higher than the corresponding negative control and contain only one peak in the melting curve.

Acknowledgments

We thank Dr. Helgi B Schiöth (supported by the Swedish Research Council) for valuable discussions regarding the manuscript. The studies were supported by the Swedish Society for Medical Research (SSMF), Svenska Läkaresällskapet, Åke Wikberg Foundation, Thurings Foundation, and Magnus Bergwalls Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2005.08.048.

References

[1] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.

[2] R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, et al., Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.
[3] R. Fredriksson, M.C. Lagerstrom, L.G. Lundin, H.B. Schioth, The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints, Mol. Pharmacol. 63 (2003) 1256–1272.
[4] R. Fredriksson, H.B. Schioth, The repertoire of G-protein coupled receptors in fully sequenced genomes, Mol. Pharmacol. (2005).
[5] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The protein kinase complement of the human genome, Science 298 (2002) 1912–1934.
[6] S. Caenepeel, G. Charydczak, S. Sudarsanam, T. Hunter, G. Manning, The mouse kinome: discovery and comparative genomics of all mouse protein kinases, Proc. Natl. Acad. Sci. USA 101 (2004) 11707–11712.
[7] F.H. Yu, W.A. Catterall, The VGL-chanome: a protein superfamily specialized for electrical signaling and ionic homeostasis, Sci STKE (2004) re15.
[8] N.H. Chen, M.E. Reith, M.W. Quick, Synaptic uptake and beyond: the sodium- and chloride-dependent neurotransmitter transporter family SLC6, Pflugers Arch. 447 (2004) 519–531.
[9] M.A. Hediger, M.F. Romero, J.B. Peng, A. Rolfs, H. Takanaga, E.A. Bruford, The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins: introduction, Pflugers Arch. 447 (2004) 465–468.
[10] E.H. Rosenberg, L.S. Almeida, T. Kleefstra, R.S. deGrauw, H.G. Yntema, N. Bahi, C. Moraine, H.H. Ropers, J.P. Fryns, T.J. deGrauw, et al., High prevalence of SLC6A8 deficiency in X-linked mental retardation, Am. J. Hum. Genet. 75 (2004) 97–105.
[11] E. Suviolahti, L.J. Oksanen, M. Ohman, R.M. Cantor, M. Ridderstrale, T. Tuomi, J. Kaprio, A. Rissanen, P. Mustajoki, P. Jousilahti, The SLC6A14 gene shows evidence of association with obesity, J. Clin. Invest. 112 (2003) 1762–1772.
[12] E. Durand, P. Boutin, D. Meyre, M.A. Charles, K. Clement, C. Dina, P. Froguel, Polymorphisms in the amino acid transporter solute carrier family 6 (neurotransmitter transporter) member 14 gene contribute to polygenic obesity in French Caucasians, Diabetes 53 (2004) 2483–2486.
[13] R.D. Blakely, L.J. De Felice, H.C. Hartzell, Molecular physiology of norepinephrine and serotonin transporters, J. Exp. Biol. 196 (1994) 263–281.
[14] H. Nelson, S. Mandiyan, N. Nelson, Cloning of the human brain GABA transporter, FEBS Lett. 269 (1990) 181–184.
[15] T. Pacholczyk, R.D. Blakely, S.G. Amara, Expression cloning of a cocaine- and antidepressant-sensitive human noradrenaline transporter, Nature 350 (1991) 350–354.
[16] B. Giros, S. el Mestikawy, N. Godinot, K. Zheng, H. Han, T. Yang-Feng, M.G. Caron, Cloning, pharmacological characterization and chromosome assignment of the human dopamine transporter, Mol. Pharmacol. 42 (1992) 383–390.
[17] G.R. Uhl, S. Kitayama, P. Gregor, E. Nanthakumar, A. Persico, S. Shimada, Neurotransmitter transporter family cDNAs in a rat midbrain library: 'orphan transporters' suggest sizable structural variations, Brain Res. Mol. Brain Res. 16 (1992) 353–359.
[18] J. Masson, C. Sagne, M. Hamon, S. El Mestikawy, Neurotransmitter transporters in the central nervous system, Pharmacol. Rev. 51 (1999) 439–464.
[19] G.S. Iyer, R. Krahe, L.A. Goodwin, N.A. Doggett, M.J. Siciliano, V.L. Funanage, R. Proujansky, Identification of a testis-expressed creatine transporter gene at 16p11.2 and confirmation of the X-linked locus to Xq28, Genomics 34 (1996) 143–146.
[20] W. Xu, L. Liu, P.A. Gorman, D. Sheer, P.C. Emson, Assignment of the human creatine transporter type 2 (SLC6A10) to chromosome band 16p11.2 by in situ hybridization, Cytogenet. Cell Genet. 76 (1997) 19.
[21] E.E. Eichler, F. Lu, Y. Shen, R. Antonacci, V. Jurecic, N.A. Doggett, R.K. Moyzis, A. Baldini, R.A. Gibbs, D.L. Nelson, Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution, Hum. Mol. Genet. 5 (1996) 899–912.
[22] H. Lill, N. Nelson, Homologies and family relationships among Na^+/Cl^- neurotransmitter transporters, Methods Enzymol. 296 (1998) 425–436.
[23] T.P. Larsson, C.G. Murray, T. Hill, R. Fredriksson, H.B. Schioth, Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery, FEBS Lett. 579 (2005) 690–698.
[24] R. Fredriksson, P.J. Hoglund, D.E. Gloriam, M.C. Lagerstrom, H.B. Schioth, Seven evolutionarily conserved human rhodopsin G protein-coupled receptors lacking close relatives, FEBS Lett. 554 (2003) 381–388.
[25] R. Fredriksson, M.C. Lagerstrom, P.J. Hoglund, H.B. Schioth, Novel human G protein-coupled receptors with long N-terminals containing GPS domains and Ser/Thr-rich regions, FEBS Lett. 531 (2002) 407–414.
[26] R. Fredriksson, D.E. Gloriam, P.J. Hoglund, M.C. Lagerstrom, H.B. Schioth, There exist at least 30 human G-protein-coupled receptors with long Ser/Thr-rich N-termini, Biochem. Biophys. Res. Commun. 301 (2003) 725–734.
[27] R. Kleta, E. Romeo, Z. Ristic, T. Ohura, C. Stuart, M. Arcos-Burgos, M.H. Dave, C.A. Wagner, S.R. Camargo, S. Inoue, et al., Mutations in SLC6A19, encoding B0AT1, cause Hartnup disorder, Nat. Genet. 36 (2004) 999–1002.
[28] H.F. Seow, S. Broer, A. Broer, C.G. Bailey, S.J. Potter, J.A. Cavanaugh, J.E. Rasko, Hartnup disorder is caused by mutations in the gene encoding the neutral amino acid transporter SLC6A19, Nat. Genet. 36 (2004) 1003–1007.
[29] S.R. Eddy, Profile hidden Markov models, Bioinformatics 14 (1998) 755–763.
[30] T. Ota, Y. Suzuki, T. Nishikawa, T. Otsuki, T. Sugiyama, R. Irie, A. Wakamatsu, K. Hayashi, H. Sato, K. Nagai, Complete sequencing and characterization of 21,243 full-length human cDNAs, Nat. Genet. 36 (2004) 40–45.
[31] A. Broer, K. Klingel, S. Kowalczuk, J.E. Rasko, J. Cavanaugh, S. Broer, Molecular cloning of mouse amino acid transport system B0, a neutral amino acid transporter related to Hartnup disorder, J. Biol. Chem. 279 (2004) 24467–24476.
[32] C.E. Burnham, B. Buerk, C. Schmidt, J.C. Bucuvalas, A liver-specific isoform of the betaine/GABA transporter in the rat: cDNA sequence and organ distribution, Biochim. Biophys. Acta 1284 (1996) 4–8.
[33] L.A. Borden, K.E. Smith, E.L. Gustafson, T.A. Branchek, R.L. Weinshank, Cloning and expression of a betaine/GABA transporter from human brain, J. Neurochem. 64 (1995) 977–984.
[34] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[35] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[36] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, J. Mol. Biol. 268 (1997) 78–94.
[37] J. Felsenstein, PHYLIP Phylogenetic inference package, Distributed by the author, Department of genetics, University of Washington, Seattle, WA (1993).
[38] S. Blair Hedges, S. Kumar, Genomic clocks and evolutionary timescales, Trends Genet. 19 (2003) 200–206.

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

[40] M.S. Springer, W.J. Murphy, E. Eizirik, S.J. O'Brien, Placental mammal diversification and the Cretaceous-Tertiary boundary, Proc. Natl. Acad. Sci. USA 100 (2003) 1056–1061.

[41] M.K. Farmer, M.J. Robbins, A.D. Medhurst, D.A. Campbell, K. Ellington, M. Duckworth, A.M. Brown, D.N. Middlemiss, G.W. Price, M.N. Pangalos, Cloning and characterization of human NTT5 and v7-3: two orphan transporters of the Na^+/Cl^--dependent neurotransmitter transporter gene family, Genomics 70 (2000) 241–252.