Glycobiology vol. 8 no. 5 pp. 517–526, 1998

The expanding β4-galactosyltransferase gene family: messages from the databanks

Neng-Wen Lo, Joel H.Shaper[1], Jonathan Pevsner[2] and Nancy L.Shaper[3]

Cell Structure and Function Laboratory, The Oncology Center and [1]Department of Pharmacology and Molecular Sciences, [2]Department of Molecular Neurobiology, Kennedy Krieger Institute and Department of Neuroscience, The Johns Hopkins University School of Medicine, Baltimore, MD 21287–8937, USA

Received on December 18, 1997; revised on January 22, 1998; accepted on January 22, 1998

[3]To whom correspondence should be addressed at: The Johns Hopkins University School of Medicine, Oncology Center, Room 1–127, 600 North Wolfe Street, Baltimore, MD 21287–8937

From a systematic search of the UniGene and dbEST databanks, using human β4-galactosyltransferase (β4GalT-I), which is recognized to function in lactose biosynthesis, as the query sequence, we have identified five additional gene family members denoted as β4GalT-II, -III, -IV, -V, and -VI. Complementary DNA clones containing the complete coding regions for each of the five human homologs were obtained or generated by a PCR-based strategy (RACE) and sequenced. Relative to β4GalT-I, the percent sequence identity at the amino acid level between the individual family members, ranges from 33% (β4GalT-VI) to 55% (β4GalT-II). The highest sequence identity between any of the homologs is between β4GalT-V and β4GalT-VI (68%). β4GalT-II is the ortholog of the chicken β4GalT-II gene, which has been demonstrated to encode an α-lactalbumin responsive β4-galactosyltransferase (Shaper *et al., J. Biol. Chem.,* 272, 31389–31399, 1997). As established by Northern analysis, β4GalT-II and -VI show the most restricted pattern of tissue expression. High steady state levels of β4GalT-II mRNA are seen only in fetal brain and adult heart, muscle, and pancreas; relatively high levels of β4GalT-VI mRNA are seen only in adult brain. When the corresponding mouse EST clone for each of the β4GalT family members was used as the hybridization probe for Northern analysis of murine mammary tissue, transcription of only the β4GalT-I gene could be detected in the lactating mammary gland. These observations support the conclusion that among the six known β4GalT family members in the mammalian genome, that have been generated through multiple gene duplication events of an ancestral gene(s), only the β4GalT-I ancestral lineage was recruited for lactose biosynthesis during the evolution of mammals.

Key words: est clones/evolution/gene duplication/lactose biosynthesis/mammary gland

Introduction

β4-Galactosyltransferase (β4GalT-I) is a constitutively expressed, *trans*-Golgi resident, type II membrane-bound glycoprotein that catalyzes the transfer of galactose to N-acetylglucosamine residues, forming the β4-N-acetyllactosamine (Galβ4-GlcNAc) or poly-β4-N-acetyllactosamine structures found in glycoconjugates (Beyer and Hill *et al.,* 1968). β4-Galactosyltransferase enzymatic activity is widely distributed in the vertebrate kingdom, in both mammals and nonmammals, including avians (Shaper *et al.*, 1997) and amphibians (unpublished observations). β4-Galactosyltransferase enzymatic activity has also been demonstrated in a subset of plants (Powell and Brew, 1974) which diverged from animals an estimated 1 billion years ago.

In mammals β4GalT-I has been recruited for a second biosynthetic function, the tissue-specific production of lactose which takes place only in the lactating mammary gland. The synthesis of lactose is carried out by the protein heterodimer assembled from β4GalT-I and the mammalian protein α-lactalbumin, a noncatalytic protein which shares a common ancestor with lysozyme (Brodbeck *et al.*, 1967). α-Lactalbumin is abundantly expressed *de novo* only in the epithelial cells of the mammary gland, beginning in mid-pregnancy and continuing throughout lactation. The notion that the β4GalT-I gene has been recruited from the nonmammalian vertebrate pool of constitutively expressed genes for lactose biosynthesis is supported by the observation that the β4GalT-I ortholog from chicken (Hill *et al.*, 1968; Shaper *et al.*, 1997) can also functionally interact with α-lactalbumin *in vitro*. Thus, the α-lactalbumin binding domain on β4GalT-I predates the rise of mammals. (Orthologs are defined as genes in different species that have evolved from a common ancestral gene; normally they retain the same function in the course of evolution. Paralogs are defined as genes related by duplication within a genome; normally, they evolve new functions even if related to the original one [Tatusov *et al.*, 1997]).

We have shown that transcription of the human and murine β4GalT-I gene in somatic cells results in two size sets of mRNAs of ~4.1 and ~3.9 kb, as a consequence of initiation at two different sets of start sites that are separated by ~200 bp. The 4.1 kb transcriptional start site is predominantly used in all somatic tissues with the notable exception of the mammary gland from mid- to late pregnant and lactating animals; in this tissue the 3.9 kb transcriptional start site is preferentially used (Harduin-Lepers *et al.*, 1993). This switch to the predominant use of the 3.9 kb start site is coincident with the cellular requirement for increased β4GalT-I enzyme levels in preparation for lactose biosynthesis. These observations, combined with a detailed promoter analysis, support a model of transcriptional regulation in which the region upstream of the 4.1 kb start site functions as a ubiquitous or housekeeping promoter for glycan biosynthesis. In contrast, the region adjacent to the 3.9 kb start site functions primarily as a mammary cell-specific promoter for lactose biosynthesis (Harduin-Lepers *et al.*, 1993; Rajput *et al.*, 1996). Based on this model, we have argued that the 3.9 kb transcriptional start site and its accompanying tissue-restricted regulatory elements have evolved in mammals to accommodate the recruited role of β4GalT-I for lactose biosynthesis (Rajput *et al.*, 1996). One prediction of this model is that the β4GalT-I ortholog in nonmammalian vertebrates, which functions exclusively in a housekeeping role (glycan biosynthesis), will exhibit a single

transcriptional start site. Consequently, we decided to characterize the β4GalT-I gene from a prototypic nonmammalian vertebrate, the chicken. The unanticipated result from this study was the demonstration that the chicken genome contains two functional, nonallelic β4GT genes (CKβ4GalT-I and CKβ4GalT-II), which encode distinct enzymatically active, α-lactalbumin responsive proteins that arose as a consequence of duplication of an ancestral gene and subsequent divergence. CKβ4GalT-I has been mapped to chicken chromosome Z in a region of evolutionary conserved synteny with the centromeric region of mouse chromosome 4 and human chromosome 9p13, where β4GalT-I had previously been mapped (Shaper *et al.*, 1986, 1990). Consequently, it is the CKβ4GalT-I ancestral lineage that has evolved into the mammalian β4GalT-I gene that is recognized to function in lactose biosynthesis, and which has been the target gene for inactivation by homologous recombination (Asano *et al.*, 1997; Lu *et al.*, 1997). In contrast, CKβ4GalT-II maps to chicken chromosome 8, in a region that is syntenic with human chromosome 1p, where a group of expressed human sequence tags (ESTs), noted to be highly similar (~55% identical) to β4GalT-I have been mapped (Shaper *et al.*, 1997).

During a systematic search of the UniGene database (Schuler *et al.*, 1996), we also found, in addition to the β4GalT-related ESTs on human 1p, four new groups of human ESTs, that were noted as being highly similar to β4GalT-I; three groups had been mapped to human chromosome 1q21–23, 3q13, and 18q11. From a search of the murine EST databank, the corresponding murine orthologs for each of the five new family members were also identified. In this study, we provide the complete coding sequence of each human β4GalT homolog and show their pattern of transcriptional expression using a panel of human somatic tissues. From an analysis of the corresponding five murine homologs, we demonstrate that only β4GalT-I is upregulated in the lactating mammary gland.

The nucleotide sequences reported in this article for human β4GalT-II, -III, -IV, -V, and -VI have been submitted to the GenBank/EMBL Data Bank with accession numbers AF038660, AF038661, AF038662, AF038663, and AF038664, respectively. The accession numbers for CKβ4GalT-I and -II are U19890 and U19889, respectively.

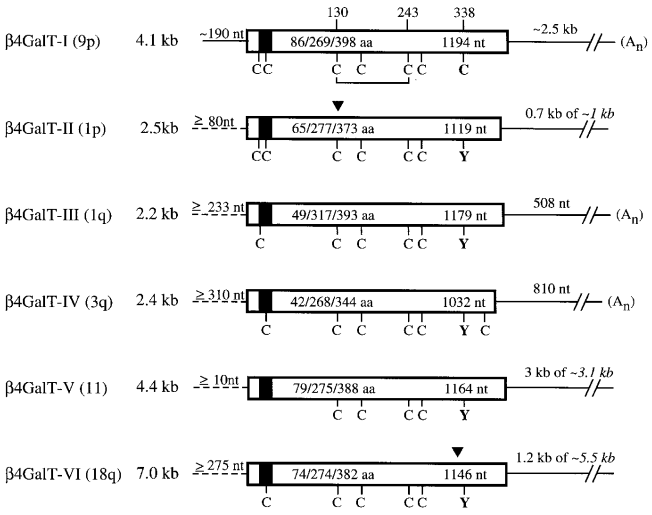

Fig. 1. Schematic representation of the human β4GalT family members. The transcript representing the gene located on human chromosome 9p13 (β4GalT-I) is shown at the top. The five additional family members (β4GalT-II through -VI) are shown with their chromosomal location and mRNA size (from Northern blot analysis) noted. The open box indicates coding sequence; the first three numbers indicate the number of amino acids in the stem region, catalytic domain and full-length coding region, respectively. The total number of nucleotides in the coding region is also shown. Since the full-length 5′-untranslated region of each homolog has not been determined, this region is depicted by a dashed line with the number of nucleotides obtained from the most 5′-clone indicated. The thin line at the right indicates the 3′-untranslated region with the number of nucleotides, available from the EST clones shown. As three of the homologs (β4GalT-II, -V, and -VI) do not contain a consensus polyadenylation signal sequence (A_n), the predicted length of the 3′-untranslated region is given in italics. The sequence of β4GalT-II and -VI that was obtained by RACE, is 5′ of the solid arrowhead. Superimposed on each mRNA is the position of the transmembrane domain (solid box) and the position of each Cys residue. The position, in β4GalT-I of the only intramolecular disulfide bond, Cys^{130} and Cys^{243} (Yadav and Brew, 1991) is indicated. As discussed, Cys^{338} (bold) in the β4GalT-I sequence is replaced by a Tyr in each family member.

Results and discussion

Search strategy used to identify five additional human β4GalT family members

We initially searched the UniGene databank (Schuler *et al.*, 1996) to identify additional β4GalT-I homologs and subsequently used the information obtained to search the dbEST databank. Examining the UniGene database first, proved to be a particularly useful search strategy as the purpose of this resource is to create a human gene catalog by clustering ESTs into groups representing distinct genes. As one gene can be represented by many sequences (e.g., alternatively spliced variants) it was decided that the presence of an identical 3′-untranslated region would define a group (unique gene). A single representative sequence from each unique gene was then mapped using one and/or two radiation hybrid panels and/or one YAC panel.

Once a group of β4GalT-related ESTs was identified, and overlapping sequences assembled, additional EST members belonging to the group were found by using the assembled nucleotide sequence as the query sequence in a search of the dbEST database using the BLASTn algorithm. This combined search revealed the presence of five additional β4GalT-I related sequence groups (genes) in the human genome, or a total of six genes when β4GalT-I is included. We have designated the family members as β4GalT-I, -II, -III, -IV, -V and -VI, where β4GalT-I represents the previously well-characterized β4GalT recognized to function in lactose biosynthesis and β4GalT-II represents the human ortholog of chicken β4GalT-II, which we described previously (Shaper *et al.*, 1997). From published studies (Shaper *et al.*, 1986) and the UniGene databank, β4GalT-I, -II, -III, -IV, and -VI have been mapped to human chromosome 9p13, 1p33–34, 1q21–23, 3q13, and 18q11, respectively (Figure 1). The chromosomal assignment for β4GalT-V has not been reported by UniGene; consequently, we used a panel of mouse/human and mouse/CHO hybrid DNAs (see *Materials and methods*) to determine that it is on human chromosome 11 (Figure 1; data not shown).

Characterization and structure of the cDNAs encoding each of the five additional human β4GalT family members

Our initial goal was to identify overlapping ESTs for each β4GalT family member (i.e., β4GalT-II, -III, etc.) that, when merged, would comprise the complete coding sequence and as much of the 5′- and 3′-untranslated regions as possible. This approach was successful for β4GalT-III, -IV, and -V, where we could account

Table I. Percent identity at the amino acid level between the vertebrate β4GalT family members

	β4GalT-I (9p)	β4GalT-II (1p)	β4GalT-III (1q)	β4GalT-IV (3q)	β4GalT-V (11)	β4GalT-VI (18q)	CKβ4GalT-I (Z)	CKβ4GalT-II (8)
β4GalT-I (9p)	100	55	50	41	38	33	69	52
β4GalT-II (1p)		100	47	41	34	34	56	72
β4GalT-III (1q)			100	47	36	36	46	46
β4GalT-IV (3q)				100	34	36	44	42
β4GalT-V (11)					100	68	35	36
β4GalT-VI (18q)						100	36	36
CKβ4GalT-I (Z)							100	53
CKβ4GalT-II (8)								100

The Genetics Computer Group GAP program was used to determine the percent sequence identity. The sequences of the CKβ4GalT-I and -II orthologs are included for comparison.

for essentially all of the full-length cDNA. For β4GalT-II and -VI the missing coding sequence was obtained using a PCR-based (RACE) strategy. All relevant EST clones used to deduce the individual coding sequences were resequenced to eliminate any errors found in the sequences deposited in the database. Next, a Northern analysis was carried out to estimate the size of the transcript encoding each new family member.

A schematic showing the structures of the transcripts for the five new β4GalT-family members, relative to β4GalT-I, is presented in Figure 1. While the coding region for each of the family members is in the range of 1–1.2 kb, the transcript sizes vary from 2.2 kb (β4GalT-III) to ~7.0 kb (β4GalT-VI). This difference in transcript size is due primarily to the length of the respective 3′-untranslated regions. Relative to the β4GalT-I mRNA, which has a 3′-untranslated region of ~2.5 kb, the 3′-untranslated regions of β4GalT-III and β4GalT-VI are 0.5 kb and ~5.5 kb, respectively. Although there is much speculation as to the role of the 3′-untranslated region in both mRNA stability and translational regulation (Decker and Parker, 1995), the significance of the vastly different sizes for individual β4GalT-family members is unknown.

Comparison of the coding regions of the six human β4GalT family members

The protein domain structure established for human β4GalT-I (398 amino acids) consists of: (1) a short NH_2-terminal cytoplasmic domain of 11 or 24 amino acids depending on the protein isoform (Shaper *et al.*, 1988; Russo *et al.*, 1990); (2) a large COOH-terminal lumenal domain (269 amino acids) containing the catalytic center, linked to a single transmembrane domain (19 amino acids) through a potentially glycosylated peptide segment of 86 amino acids, termed the stem region. The catalytic domain can be further subdivided into two distinct structure/function subdomains. (i) The NH_2-terminal region of the catalytic domain contains a 113 amino acid loop formed by the only intramolecular disulfide bond present in the protein, between Cys^{130} and Cys^{243} (see schematic in Figure 1). This loop plus adjacent sequence in the stem region (the stem region is defined as the amino acid sequence between the transmembrane domain and Cys^{130}) is involved in α-lactalbumin binding as established by protection studies (Yadav and Brew, 1990) and antibody blocking studies (Ulrich *et al.*, 1986; Russo, 1990). (ii) The COOH-terminal 157 amino acid segment contains two polypeptides, in the vicinity of Cys^{338} (Figures 1 and 2), that can be affinity-labeled with UDP-Gal analogues (Aoki *et al.*, 1990; Yadav and Brew, 1990) or have been implicated in substrate binding by site directed mutagenesis (Aoki *et al.*, 1990).

A global alignment of β4GalT-I and the five additional β4GalT homologs is shown in Figure 2. The percentage amino acid sequence identity between each homolog is summarized in Table I. As seen in the schematic (Figure 1), each homolog encodes a type-II transmembrane protein with sizes ranging from 344 (β4GalT-IV) to 393 amino acids (β4GalT-III). However, the size of the respective catalytic domain is more tightly clustered between 268 (β4GalT-IV) and 277 (β4GalT-II) amino acids; β4GalT-III which has a catalytic domain of 317 amino acids, due to a COOH-terminal extension of ~42 amino acids, stands out as the only exception to this general pattern (Figure 2). The main difference in protein domain structure is in the lengths of the respective stem regions (the region between the transmembrane domain and Cys^{130} in the β4GalT-I sequence) which range from 42 (β4GalT-IV) to 86 (β4GalT-I) amino acids.

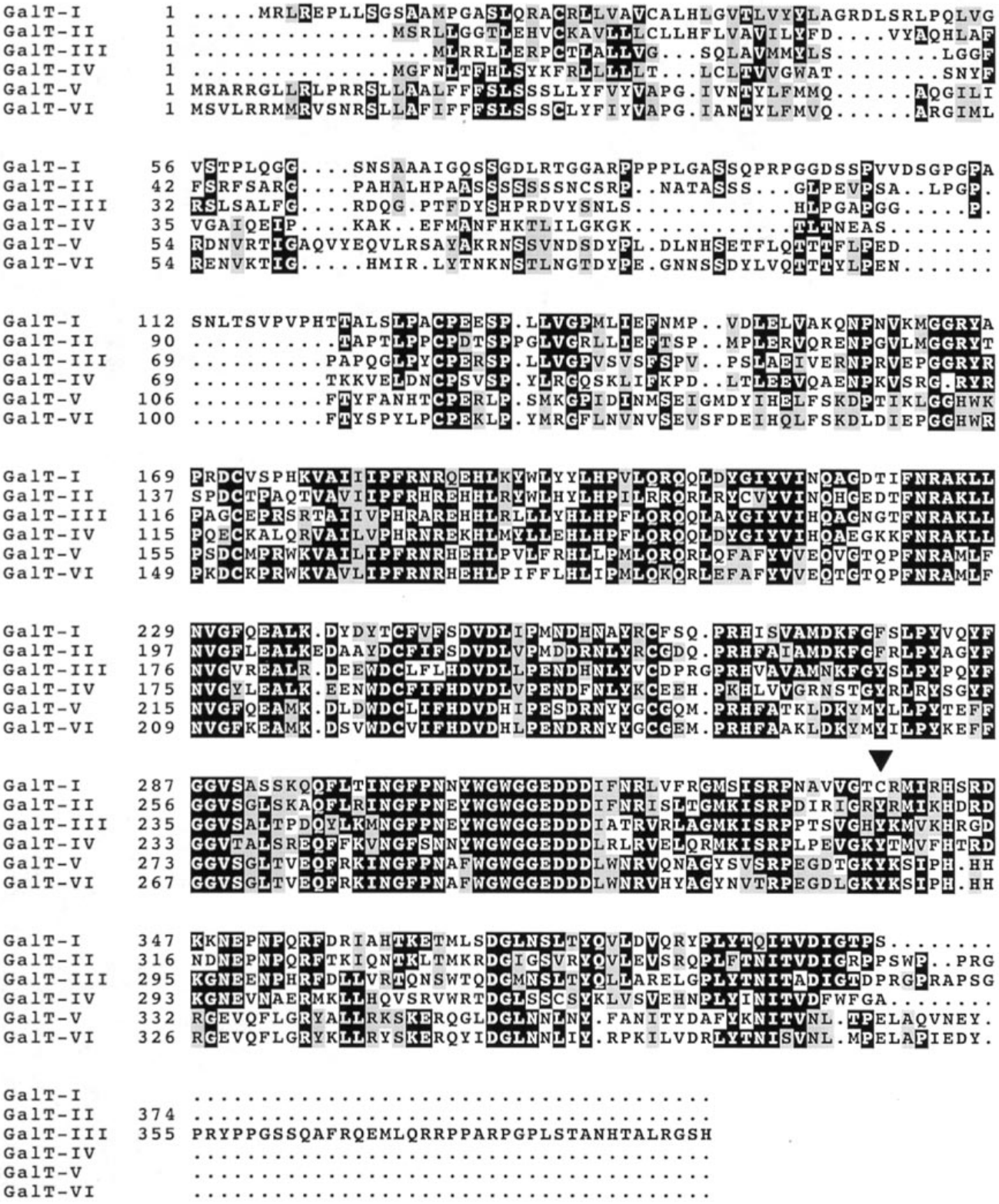

Fig. 2. Amino acid sequence alignment of the human β4GalT family members using the ClustalW program. Black boxes indicate identical residues in all six proteins; gray boxes indicate conserved residues. The position of the Cys to Tyr substitution is indicated by the arrowhead.

As summarized in Table I, the percent sequence identity at the amino acid level between the individual human homologs, relative to human β4GalT-I, ranges from 33% (β4GalT-VI) to 55% (β4GalT-II). The highest sequence identity between any of the human homologs is between β4GalT-V and -VI (68%). When CKβ4GalT-I and CKβ4GalT-II are included in the comparison,

		Cytoplasmic Domain	Transmembrane Domain	Stem Region
β4GalT-I	9p	MRLREPLLSGSAAMPGASLQRACR	LLVAVCALHLGVTLVYYLA	GRDLSRL..
β4GalT-II	1p	MSRLLGGTLERVCK	AVLLLCLLHFLVAVILYF	DVYAQHL..
β4GalT-III	1q	MLRRLLERP	CTLALLVGSQLAVMMYLSL	GGFRSLS..
β4GalT-IV	3q	MGFNLTFHLSYKFR	LLLLFTLCLTVVGWATSNYFV	GAIQEIP..
β4GalT-V	11	MRARRGLLRLPRR	SLLAALFFFSLSSSLLYFVYVA	PGIVNTY..
β4GalT-VI	18q	MSVLRRMMRVSNR	SLLAFIFFFSLSSSCLYFIYVA	PGIANTY..

Fig. 3. Alignment of the NH_2-terminal cytoplasmic domain, transmembrane domain and first seven residues of the stem region of the β4GalT family members. The putative transmembrane domain was identified using the TMpred program (Hofmann and Stoffel, 1993). Although it had been reported that the cytoplasmic domain of the human sequence lacks the Ser residue at amino acid 11, when the human cDNA was resequenced, we found that the trinucleotide encoding this residue was present (Shaper *et al.*, 1997).

they exhibit 69 and 72% identity with their corresponding human orthologs (also see Shaper *et al.*, 1997).

From an inspection of Figure 2, it is clear that the respective catalytic domains of the β4GalT-family members are highly conserved. The structural domains that are least conserved are the stem domain and the NH_2-terminal region of the cytoplasmic domain (Figures 2 and 3). Of particular note are the presence or lack thereof, of the Cys residues found in human β4GalT-I. Only the first four Cys residues in the lumenal/catalytic domain, including the two involved in the single intramolecular disulfide bond (Cys^{130} and Cys^{243} in Figures 1 and 2; Yadav and Brew, 1991), are conserved in each family member. The Cys^{338} residue is found only in β4GalT-I (and CKβ4GalT-I); in β4GalT-II (and CKβ4GalT-II) this Cys residue is replaced by Tyr. As discussed previously (Shaper *et al.*, 1997), this fortuitous Cys to Tyr replacement is a useful marker to follow the evolutionary gene lineage of CKβ4GalT-I and CKβ4GalT-II in the human and mouse genomes. Based on this criterion, it would appear that earliest ancestor of the vertebrate β4GalT gene family had a Tyr in this position.

A multiple sequence alignment of the NH_2-terminal region, including the cytoplasmic and transmembrane domain, is presented in Figure 3. The lengths of the cytoplasmic domains range from 9 (β4GalT-III) to 24 amino acids (β4GalT-I) while the lengths of the putative transmembrane domains range from 18 to 22 amino acids. The transmembrane domain and perhaps the flanking amino acids in the cytoplasmic domain have been demonstrated to contain the β4GalT-I *trans*-Golgi retention signal (reviewed by Colley, 1997). In this context it is interesting that the amino acid sequence of the respective transmembrane domains are highly divergent. It will be of interest to determine the sub-Golgi localization of each new β4GalT homolog to determine if the corresponding region(s) are also responsible for Golgi retention.

Phylogenetic analysis of the vertebrate β4GalT family members

An inferred phylogenetic tree (cladogram) was constructed to analyze the evolutionary relationships between the six human and two chicken β4GalT homologues (Figure 4). (A cladogram is a diagram which depicts a hypothetical branching sequence of lineages leading to the taxa under consideration. A clade, from the Greek "klados," meaning branch or twig, is a group of organisms which includes their most recent common ancestor and all of its descendants. For a detailed overview of phylogeny, refer to the "Tree of Life" web site at http://phylogeny.arizona.edu/tree/phylogeny.html). To construct the tree, the eight sequences were multiply aligned and character positions containing any gaps were eliminated; for each protein, 193 amino acid residues were aligned. Parsimony analysis was used to construct a tree that required the minimal number of evolutionary changes to account for the differences among the six human and two chicken β4GalT family members at each amino acid position. (Parsimony refers to a rule used to choose among possible cladograms, which states that the cladogram implying the least number of changes in character states is the best.) The tree is unrooted because no ancestral β4GalT is known to define an outgroup. (An outgroup, in a cladistic analysis, is a taxon used to help resolve the states of characters, and which is hypothesized to be less closely related to each of the taxa under consideration than any are to each other.) In an unrooted tree such as this there is no root node and branch lengths specify relationships among the β4GalTs without defining a primordial evolutionary path (reviewed in Li and Grauer, 1991).

To gain a statistical measure of confidence in the tree, we performed a bootstrap analysis. A total of 100 trees were generated from the initial data set, and the percentage of trees containing a particular clade was measured. (A clade is a group of β4GalT family members that contains a common ancestor that is not shared by any family member outside the group.) Bootstrap values >70% are associated with statistical significance at the $P < 0.05$ level (Hillis and Bull, 1993).

The cladogram indicates that the eight vertebrate β4GalT family members cluster into four groups: β4GalT-I (human and chicken); β4GalT-II (human and chicken); β4GalT-III and -IV; and β4GalT-V and -VI. Three of these groupings had high bootstrap percentages (92%, 98%, 100%), indicating that they are likely to represent authentic clades. However, the β4GalT-III and -IV clade was reproduced in only 67% of the samplings, indicating that while these two proteins could represent an authentic cluster, in 33% of the data samplings this particular clade was disrupted by the positioning of one of these proteins in a different region of the cladogram.

This cladogram highlights the ancestral lineage between human and chicken β4GalT-I proteins, as well as between the β4GalT-II proteins. As previously discussed, the evolution of the β4GalT-I and β4GalT-II proteins must have occurred as a gene duplication event prior to the divergence of human and chicken lineages 250 million years ago (Shaper *et al.*, 1997). The subsequent speciation event of humans and chickens and subsequent divergence has resulted in β4GalT-I and CKβ4GalT-I protein orthologs with 69% amino acid identity (Table I). Interestingly, this degree of amino acid identity is similar to that

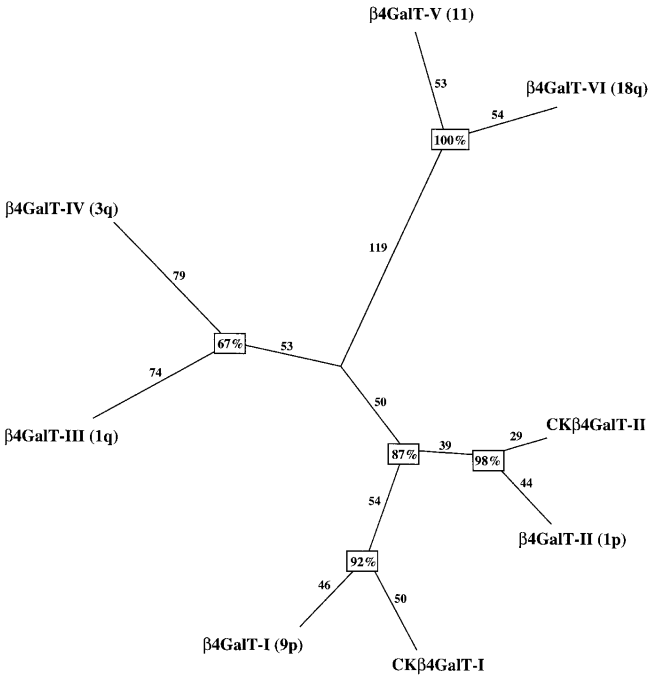

Fig. 4. Phylogenetic analysis of the human and chicken β4GalT family members. The tree represents an unrooted cladogram. Branch length values are indicated and are additive. Bootstrap values are indicated by the boxed numbers. These percentages derive from sampling 100 trees to obtain confidence values for the groupings of particular clades. The name of each human homolog is indicated as is its chromosomal position.

observed with other known human and chicken Golgi-resident, terminal glycosyltransferases such as the human and chicken α1,3-fucosyltransferase (63% amino acid identity; accession numbers M65030 and U73678, respectively) or the human and chicken α-2,3-sialyltransferase (67% amino acid identity; accession numbers L29555 and X80503, respectively).

The branch lengths of the cladogram indicate inferred evolutionary distance and reflect the number of reconstructed amino acid changes (i.e., substitutions) on the branch. Thus, for example, human β4GalT-I and CKβ4GalT-I are separated by branch lengths of 46 and 50 amino acid residues, reflecting the number of amino acid substitutions along the length of each protein required by parsimony analysis to account for their sequence divergence from a common ancestor. The branch lengths of the four major groups in the cladogram (Figure 4; β4GalT-I,-II, -III/-IV, and -V/-VI) suggest that these groups are approximately equidistant. The branch lengths of β4GalT-V and -VI, connecting these two proteins to other members of the cladogram, are somewhat longer, suggesting that for a constant rate of nucleotide substitution, these genes are ancestral in the β4GalT family.

Phylogenetic analysis of the invertebrate and vertebrate β4GalT family members

To further characterize the β4GalT gene family, we performed BLAST searches of GenBank databases to identify homologs in other species. In addition to the six human and two chicken sequences, we identified one sequence from mouse, and one from bovine which are the corresponding β4GalT-I orthologs. Additionally two proteins from the nematode *C.elegans* and two from the snail *L.stagnalis*, were also detected giving a total of 14 sequences. One of the snail proteins (*L.stagnalis*-2) has been identified as a UDP-GlcNAc:GlcNAc β4-N-acetylglucosaminyltransferase (Bakker *et al.*, 1994). The identification of the *L.stagnalis*-1 protein and the proteins encoded by the two *C.elegans* genes has not been reported.

The 14 β4GalT related sequences were multiply aligned and all positions with gaps were eliminated, resulting in an alignment of proteins across 134 amino acid residues. The relation of these proteins was evaluated by constructing a cladogram by parsimony analysis (Figure 5). This tree thus contains additional β4GalT members, but the portions of aligned proteins are smaller (less informative) than the regions of aligned proteins described in Figure 4. All proteins analyzed in the cladogram shared statistically significant amino acid identity, as determined by their Z-scores (data not shown; see Materials and methods). The cladogram consists of the same four overall groupings as in the previous tree, as well as two additional groups. The first cluster includes the murine and bovine β4GalT-I orthologs closely related to human β4GalT-I. The clusters of β4GalT-II and β4GalT-III/-IV are similar to those described in Figure 4. The fourth cluster of human β4GalT-V and -VI is joined by a *C.elegans* protein which we denote as *C.elegans*-2. Two additional features are present in the tree. The two snail proteins form a cluster and are 71% identical to each other, over the 134 amino acid region that was compared. They share 31–35% amino acid identity to human β4GalT-III,-IV, -V, and -VI. Separately, a hypothetical nematode protein that we designate *C.elegans*-1 has between 21% and 27% identity to all 13 other proteins in the cladogram. Thus, while *C.elegans*-1 is homologous to other β4GalT family members, its ortholog was not identified in BLAST searches of the database of human or mouse expressed sequence tags.

Do the β4GalT homologs show tissue restricted expression?

Northern blot analysis, in combination with quantitation by means of phosphorimaging, was performed to determine in which tissue type(s) each homolog is expressed, and to estimate the respective steady state mRNA levels relative to β4GalT-I. The results of this analysis are shown in Figure 6. β4GalT-I is constitutively expressed in all human tissues examined with the exception of both fetal and adult brain, where steady state mRNA levels are reduced by ~80%. This pattern of expression for β4GalT-I observed in human tissues is consistent with results obtained in murine tissues (Harduin-Lepers *et al.*, 1993).

β4GalT-III is also constitutively expressed at comparable levels to β4GalT-I in the human tissues examined; however, in contrast to β4GalT-I, β4GalT-III is also expressed in high levels in the fetal brain and in somewhat lower levels in the adult brain. A somewhat similar pattern is also exhibited by β4GalT-V, although overall expression levels appear to be lower. β4GalT-IV also appears to be widely expressed at low levels, although the adult brain, lung, and liver show only a very weak signal. β4GalT-II and -VI show the most restricted pattern of tissue expression. High steady state levels of β4GalT-II mRNA are seen only in fetal brain and adult heart, muscle, and pancreas. Relatively high steady state levels of β4GalT-VI mRNA are seen only in adult brain.

Are any of the additional β4GalT homologs expressed in the murine mammary gland during lactation?

During the second half of pregnancy, β4GalT-I enzyme levels in the mammary epithelial cell rise ~50-fold in preparation for the

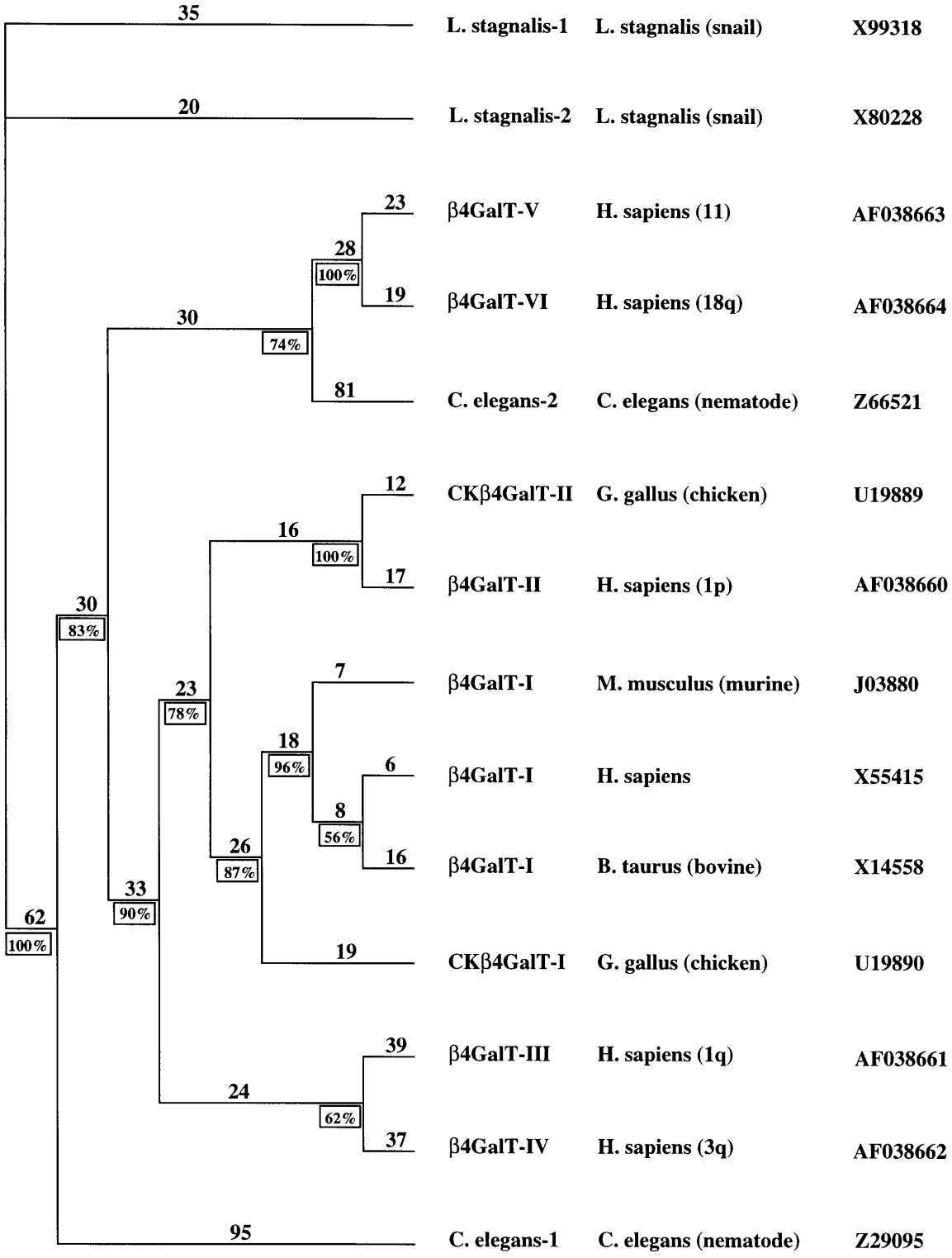

Fig. 5. Phylogenetic analysis of the β4GalT family of proteins. Fourteen full-length sequences were identified. The tree represents an unrooted cladogram. Branch lengths are indicated and are additive. Bootstrap values are indicated by the boxed numbers. The names of the proteins are indicated with the species and DNA accession numbers.

production of lactose (Turkington *et al.*, 1968; Palmiter, 1969). Mechanistically, this increase is achieved in part, by a switch from the use of the 4.1- to the 3.9 kb transcriptional start site, which is governed by a stronger promoter that is operative primarily in the mammary gland during lactation (Rajput *et al.*, 1996). As discussed in the Introduction, we have argued that the 3.9 transcriptional start site and its accompanying tissue-restricted regulatory elements have been introduced into the ancestral β4GalT-I gene lineage during the evolution of mammals to accommodate the recruited role of β4GalT-I for lactose biosyn-

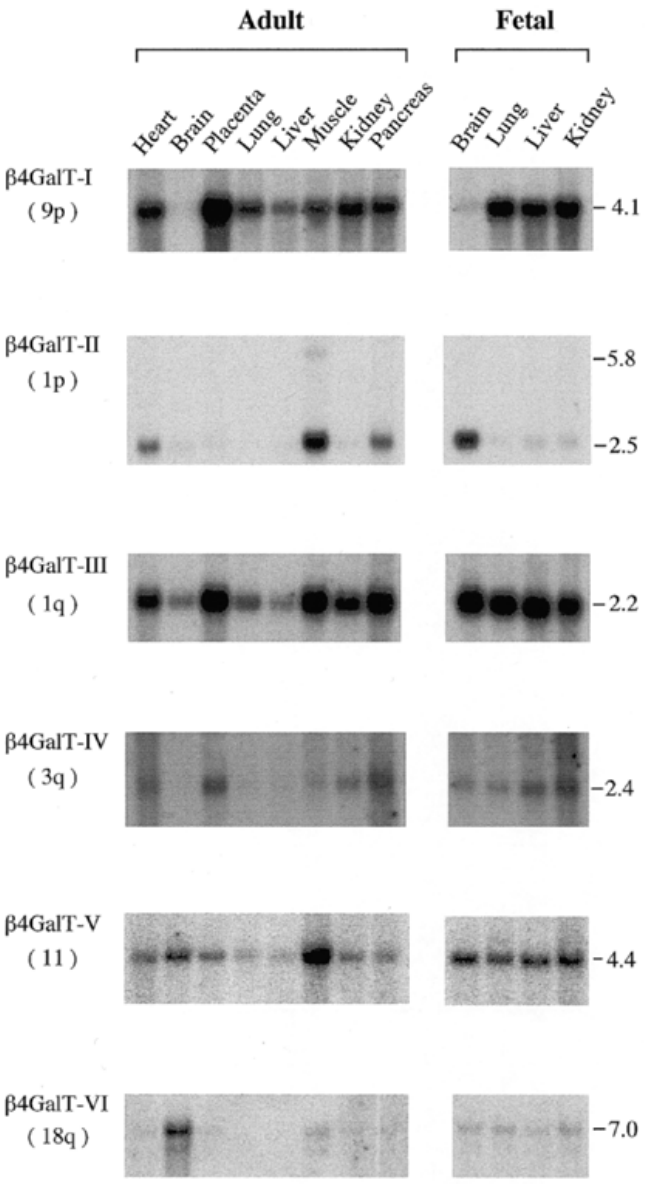

Fig. 6. Expression levels of the β4GalT family members in various human tissues. Human multiple tissue Northern blots, containing 2 μg of poly $(A)^+$ RNA isolated from each tissue, were hybridized using an ~800 bp probe corresponding to each human homolog. All probes were labeled to the same specific activity and exposure times (4 days) were identical.

thesis (Rajput *et al.*, 1996). Since multiple transcription factor binding sites, including that of the tissue restricted transcription factor AP2, are involved in the expression of this mRNA, we would anticipate that the assembly of this tissue specific promoter would have occurred only once during evolution.

To carry out this analysis, we choose to use the murine system because of the relative ease in obtaining the appropriate tissue. RNA was obtained from the murine lactating mammary gland and a murine clone, corresponding to each human β4GalT homolog, was identified using the strategy described in *Materials and methods*. A PCR fragment of ~800 bp was subsequently generated from the appropriate mouse EST clone for use as the hybridization probe for Northern analysis. As summarized in Table II, expression of only the β4GalT-I gene could be detected in the lactating mammary gland. This result is particularly interesting in considering the biological function of the mammalian β4GalT-II gene, which is the ortholog of the CKβ4GalT-II gene, previously demonstrated to encode a functional α-lactalbumin responsive β4-galactosyltransferase (Shaper *et al.*, 1997). In the absence of mammary gland promoter element(s) that are operative during lactation, the mammalian β4GalT-II gene is not expressed in sufficient levels during lactation to contribute significantly to lactose biosynthesis. This conclusion is consistent with recently reported studies in which murine β4GalT-I was inactivated by homologous recombination (Asano *et al.*, 1997; Lu *et al.*, 1997). One of the main problems observed in null mothers was the inability to produce lactose. In summary, these results support the conclusion that among the six known β4GalT family members in the mammalian genome, that have been generated through multiple gene duplication events of an ancestral gene(s), only the β4GalT-I ancestral lineage was recruited for lactose biosynthesis during the evolution of mammals.

Table II. Expression of the β4GalT family members in the murine lactating mammary gland

β4GalT-I	+
β4GalT-II	-
β4GalT–III	-
β4GalT-IV	-
β4GalT-V	-
β4GalT-VI	-

The levels of expression were determined by Northern blot analysis using a probe derived from the corresponding mouse homolog. A Northern blot showing β4GalT-I expression in the lactating mammary gland is shown in Figure 1 in Harduin-Lepers *et al.* (1993).

Materials and methods

cDNA clones encoding the β4GalT family members

cDNA clones were obtained from Genome Systems, Inc. (St. Louis, MO) or ATCC (Rockville, MD) and are designated by accession number. Overlapping clones were chosen for sequencing that contained the protein coding sequence. Of the 10 EST sequences encoding β4GalT-II, W07207, R01345, and AA453005 were sequenced. Of the 53 EST sequences encoding β4GalT-III, H30715, AA055202, and W88517 were sequenced. Of the 29 EST sequences encoding β4GalT-IV, AA101851, and AA046963 were sequenced. Of the 40 EST sequences encoding β4GalT-V, AA243575, AA293458, AA476439, and AA223560 were sequenced. Of the three EST sequences encoding β4GalT-VI, R19559 was sequenced.

Identification of the murine β4GalT orthologs

The BLASTn program was used to search the dbEST database for murine orthologs, using the nucleotide sequence from the coding region of each of the five human β4GalT family members. A mouse sequence producing a high-scoring segment pair was then aligned, using the MacVector DNA pustell Matrix program, to the query human β4GalT sequence as well as to each β4GalT family member. The dbEST sequence was considered to be the candidate mouse ortholog if identity to only the original query sequence was >90% .

Chromosomal assignment of β4GalT-V

The National Institute of General Medical Science (NIGMS) monochromosomal panel was obtained from the Core Facility of

Johns Hopkins. This panel of 24 DNA samples consists of either human/mouse or human/CHO DNA hybrids with a single human chromosome present in each hybrid (Drwinga *et al.*, 1993). Each hybrid DNA (100 ng) plus control mouse and CHO DNA was transferred to a Nytran membrane using a slot blot apparatus and hybridized with a probe representing ~800 bp of the 3′-untranslated region of the cDNA encoding β4GalT-V. The probe used was a PCR fragment generated using the following forward and reverse primers, respectively: 5′-GAATGTACGTTTGCTTTACCCCA-3′; 5′-GCTACGCTCAATGCCATCGTC-3′; the target was human genomic DNA. After washing at high stringency, the only slot that showed positive hybridization contained DNA from human chromosome 11.

Northern blot analysis and probes

Human multiple tissue Northern blots, obtained from Clontech (Palo Alto, CA), contained 2 μg poly $(A)^+$ RNA isolated from each of the designated tissues. Duplicate blots for each set were obtained to avoid the necessity of stripping any one blot. [^{32}P]-Labeled cDNA probes of similar specific activity were used for hybridization. Probes were generated by PCR using gene specific primers or by digestion of the appropriate clone with restriction enzymes. The PCR primers (e.g., 9pF is name given to the forward primer used for amplification of the gene on chromosome 9p), and designated target DNA (either genomic DNA or an EST clone) were as follows: 9pF 5′-GTCAGGATCTGCCGGCACGAAAG-3′ and 9pR 5′-CTTTCTGTCCGCAGATCCTGAC-3′, human fibroblast genomic DNA; 1pF 5′-AGTTTCAGAACCACTTTGGG-3′ and 1pR M13F primer, W07207; 1qF 5′-GCAAGATGGGATGAACTCACT-3′ and 1qR M13F primer, W88517; 3qF 5′-TGACCCTGGATCTTTTGGTGAT-3′ and 3qR 5′-TGTATTCTCTGGTGGGCATCA-3′, human fibroblast genomic DNA; 18qF 5′-TCATGCCAGAGTTAGCTCCA-3′ and 18qR M13F. AA243575 was digested with *Not*I and *Eco*RI to obtain a probe for the gene on chromosome 11. Blots were washed at high stringency and exposed to Kodak XAR-5 film for 4 days.

Northern blots containing 3 μg of poly $(A)^+$ RNA isolated from murine lactating mammary glands were prepared as described (Harduin-Lepers *et al.*, 1993). [^{32}P]-Labeled cDNA probes of similar specific activity were used for hybridization. Probes were generated by PCR using gene specific primers or by digestion of a cDNA clone with restriction enzymes. The PCR primers (e.g., 1pF is name given to the forward primer used to amplify the mouse β4GalT-II homolog) and designated target DNA (a murine EST clone) were as follows: 1pF 5′-GGCGAGGATGATGACATCTT-3′ and 1pR 5′-AAGCATGAGGGGTCTCCAAA-3′, W77594; 1qF 5′-AGGAGCAGGGCTGGACCCCA-3′ and 1qR M13F, W34108; 3qF 5′-CGGCATCTATATCATCCACC-3′ and 3qR 5′-CTTCACAGCCATGATTCAAA-3′, AA111257; 11F 5′-CTACCTCTTCATGCTGCAGG-3 and 11R 5′-CCACAAGTCGTCATCTTCTC-3′, AA013728; 18qF 5′-TCTATTCCTCATCACCATCG-3′ and 18qR 5′-CCAACAATTTGAACACATTT-3′, AA414080. A 780 bp *Eco*RI fragment, derived from the 3′-untranslated region of the murine cDNA clone MGT-1 (Shaper *et al.*, 1988), was used for the murine β4GalT-I equivalent. Blots were washed at high stringency and exposed to Kodak XAR-5 film for 4 days.

RACE

The 5′-end of the transcript encoding β4GalT-II and -VI was obtained using the Marathon cDNA Amplification Kit from Clontech (Palo Alto, CA) following the manufacturer's instructions. Marathon ready cDNA from human fetal brain and human adult brain (Clontech) was used as the starting material for β4GalT-II and -VI, respectively. The gene specific primer for β4GalT-II was 5′-TAAAGGGGATGATGACCGCCAC-3′ and the nested specific primer was 5′-TGAACTCGATCAGCAGTCTG-3′. The gene specific primer for β4GalT-VI was 5′-CCTAAGTCTCCCTCTGGTCTGGTTAC-3′ and the nested specific primer was 5‘-CTCTGTTCCAAAGGTCATCATCTTCTCC-3′. Fragments were subcloned into the TA cloning vector (Invitrogen, San Diego, CA) and sequenced.

Computer analyses

The UniGene database can be found at http://www.ncbi.nlm.nih.gov/. The basic local alignment search tool (blastn and tblastn algorithms) was used to search the GenBank dbEST database. Sequence comparisons were performed using MacVector, AssemblyLIGN (International Biotechnologies, Inc., New Haven, CT) and the Genetics Computer Group (Madison, WI) GAP program. Statistical significance for the relatedness of two proteins was determined with GAP by generating Z scores. Z scores were obtained by measuring the quality score between two proteins, subtracting the mean quality score obtained from comparisons with 50 randomized shuffles of one protein, and dividing this value by the standard deviation of those 50 scores. Z scores above 3 are considered statistically significant. The percent sequence identities were determined using the Genetics Computer Group GAP program. Multiple sequence alignments were performed using ClustalW 1.7; BOXSHADE was used to format Figure 2. The transmembrane domains were determined using TMpred (Hofmann and Stoffel, 1993). The latter three programs can be accessed via the following web site: http://www.public.iastate.edu/~pedro/research_tools.html. The Phylogenetic Analysis Using Parsimony (PAUP) program, prerelease version 4.0d60, was used to generate phylogenetic trees and was generously provided by Dr. David Swafford of the Smithsonian Institute.

Acknowledgments

This work was supported in part by National Institutes of Health Grant CA45799 (to J.H.S.) and March of Dimes Grant 5-FY96–1177 (to J.P.). Neng-Wen Lo is a postdoctoral fellow supported by Grant 960188 (to J.H.S.) from the Mizutani Foundation for Glycoscience.

Abbreviations

β4GalT-I refers to the α-lactalbumin responsive, UDP-galactose:N-acetylglucosamine β4-galactosyltransferase (β1,4-galactosyltransferase (EC 2.4.1.38)) that has been mapped to human chromosome 9p13, and the centromeric region of mouse chromosome 4, respectively, whereas CKβ4GalT-I refers to the chicken ortholog that has been mapped to chromosome Z; CKβ4GalT-II refers to the α-lactalbumin responsive, β1,4-galactosyltransferase that has been mapped to chicken chromosome 8, whereas β4GalT-II refers to the human ortholog mapped to human chromosome 1p, or the mouse ortholog; β4GalT-III, β4GalT-IV, β4GalT-V, and β4GalT-VI denote the human β4-galactosyltransferase homologs that have been mapped to human chromosome 1q21–23, 3q13, 11 and 18q11, respectively; aa, amino acid; BLAST, basic local alignment search tool; bp, base pair(s); EST(s), expressed sequence tag(s); nt, nucleotide(s);

PCR, polymerase chain reaction; RACE, rapid amplification of cDNA ends; YAC, yeast artificial chromosome.

Note added in proof

Two human β4-galactosyltransferase genes, designated β4GalT-2 and β4GalT-3 have been independently reported (Alemida *et al.*, J. Biol. Chem. **272,** 31979–31991, 1997). β4GalT-2 is the human ortholog of the α-lactalbumin responsive, chicken β4-galactosyltransferase, designated CKβ4GalT-II, previously reported (Shaper *et al.*, 1997). In this article we refer to this human ortholog as β4GalT-II. The β4GalT-3 gene, which fortuitously corresponds to the gene we had designated β4GalT-III, encodes an α-lactalbumin nonresponsive β4-galactosyltransferase activity.

It is important to note that the amino acid sequence presented by Almeida *et al.*, which comprises the NH_2-terminal cytoplasmic domain of the β4GalT-2/β4GalT-II protein, differs significantly from the sequence that we have presented. We have subsequently resequenced this region from both human cDNA and genomic DNA and also have sequenced the cDNA of the murine β4GalT-II ortholog. The data obtained from all three DNA sources confirms that the nucleotide sequence, as reported in our study, is correct. Instead of six G residues at nt positions 90–95 (see Figure 2 in Almeida *et al.*,) there are seven G residues. The insertion of the additional G residue alters the reading frame such that the ATG at nt 76–78 (see Figure 2 in Almeida *et al.*) becomes the initiating Met. This change in nucleotide sequence means that an intron is not positioned within the coding sequence in the NH_2-terminal region as indicated by Almeida *et al.*, (see Figure 10). Instead, this intron is positioned in the 5′-untranslated region, ~45 nt upstream of the initiating ATG. The position of this intron in the 5′-untranslated region of the β4GalT-II gene results in a gene structure that is identical to the chicken ortholog, CKβ4GalT-II (Shaper *et al.*, 1997).

A fourth human β4-galactosyltransferase gene has been reported (Sato *et al.*, *Proc. Natl. Acad. Sci.*, **95,** 472–477, 1998). It has been expressed as a protein-A fusion protein and demonstrated to encode a β4-galactosyltransferase activity that is not responsive to α-lactalbumin. Based on sequence, this β4GalT gene corresponds to the gene designated β4GalT-V in this article.

This recent data is interesting in the context of the cladogram presented in Figure 4. Based on the analysis of expressed recombinant proteins, at least one human β4GalT family member from each of the four clades has been demonstrated to encode a UDP-galactose:N-acetylglucosamine β4-galactosyltransferase activity by direct assay.

References

Aoki,D., Appert,H.E., Johnson,D., Wong,S.S. and Fukuda,M.N. (1990) Analysis of the substrate binding sites of human galactosyltransferase by protein engineering. *EMBO J.*, **9,** 3171–8.

Asano,M., Furukawa,K., Kido,M., Matsumoto,S., Umesaki,Y., Kochibe,N. and Iwakura,Y. (1997) Growth retardation and early death of β1,4-galactosyltransferase knockout mice with augmented proliferation and abnormal differentiation of epithelial cells. *EMBO J.*, **16,** 1850–7.

Bakker,H., Agterberg,M., Van Tetering,A., Koeleman,C.A., Van den Eijnden,D.H. and Van Die,I. (1994) A *Lymnaea stagnalis* gene, with sequence similarity to that of mammalian β1→4-galactosyltransferases, encodes a novel UDP-GlcNAc:GlcNAc β-R β 1→4-N-acetylglucosaminyltransferase. *J. Biol. Chem.*, **269,** 30326–33.

Beyer,T.A. and Hill,R.L. (1968) Glycosylation pathway in the biosynthesis of nonreducing terminal sequences in oligosaccharides of glycoproteins. In Horowitz,M. (ed.), *The Glycoconjugates.* Vol. III, Academic Press, New York, pp. 25–45.

Brodbeck,V., Denton,W.L., Tanahashi,N. and Ebner,K.E. (1967) The isolation and identification of the β protein of lactose synthetase as α-lactalbumin. *J. Biol. Chem.*, **242,** 1391–1397.

Colley,K.J. (1997) Golgi localization of glycosyltransferases: more questions than answers. *Glycobiology*, **7,** 1–13.

Decker,C.J. and Parker,R. (1995) Diversity of cytoplasmic functions for the 3′ untranslated region of eukaryotic transcripts. *Curr. Opinion Cell Biol.*, **7,** 386–392.

Drwinga,H.L., Toji,L.H., Kim,C.H., Greene,A.E. and Mulivor,R.A. (1993) NIGMS Human/rodent somatic cell hybrid mapping panels 1 and 2. *Genomics*, **16,** 311–314.

Harduin-Lepers,A., Shaper,J.H. and Shaper,N.L. (1993) Characterization of two cis-regulatory regions in the murine β1,4-galactosyltransferase gene: evidence for a negative regulatory element that controls initiation at the proximal site. *J. Biol. Chem.*, **268,** 14348–14359.

Hill,R.L., Brew,K., Vanaman,T.C., Trayer,I.P. and Mattock,P. (1968) The structure, function and evolution of α-lactalbumin. *Brookhaven Symp. Biol.*, **21,** 139–154.

Hillis,D.M. and Bull,J.J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.*, **42,** 182–192.

Hofmann,K. and StoffelW. (1993) Tmbase—a database of membrane spanning protein segments. *Biol. Chem. Hoppe-Seyler,* **347,** 166.

Li,W.-H. and Graur,D. (1991) *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Lu,Q., Hasty,P. and Shur,B.D. (1997) Targeted mutation in β1,4-galactosyltransferase leads to pituitary insufficiency and neonatal lethality. *Dev. Biol.*, **181,** 257–267.

Palmiter,R.D. (1969) Hormonal induction and regulation of lactose synthetase in mouse mammary gland. *Biochem. J.*, **113,** 409–417.

Powell,J.T. and Brew,K. (1974) Glycosyltransferases in the Golgi membranes of onion stem. *Biochem. J.*, **142,** 203–209.

Rajput,B., Shaper,N.L. and Shaper,J.H. (1996) Transcriptional regulation of murine β1,4-galactosyltransferase in somatic cells: analysis of a gene that serves both a housekeeping and a mammary gland-specific function. *J. Biol. Chem.*, **271,** 5131–5142.

Russo,R.N. (1990) Two forms of β1,4-galactosyltransferase, Ph.D. thesis. Johns Hopkins University.

Russo,R.N., Shaper,N.L. and Shaper,J.H. (1990) Bovine β1,4-galactosyltransferase: two sets of mRNA transcripts encode two forms of the protein with different amino terminal domains: *in vitro* translation experiments demonstrate that both the short and the long forms of the enzyme are type II membrane-bound glycoproteins. *J. Biol. Chem.*, **265,** 3324–3331.

Schuler,G.D. *et al.* (1996) A gene map of the human genome. *Science*, **274,** 540–546.

Shaper,N.L., Shaper,J.H., Bertness,V., Chang,H., Kirsch,I.R. and Hollis,G.F. (1986) The human galactosyltransferase gene is on chromosome 9 at band p13. *Somatic Cell Mol. Genet.*, **12,** 633–636.

Shaper,N.L., Hollis,G.F., Douglas,J.G., Kirsch,I.R. and Shaper,J.H. (1988) Characterization of the full-length cDNA for murine β1,4-galactosyltransferase: novel features at the 5′ end predict two translational start sites at two in-frame AUGs. *J. Biol. Chem.*, **263,** 10420–10428.

Shaper,N.L., Shaper,J.H., Peyser,M. and Kozak,C.A. (1990) Localization of the gene for β1,4-galactosyltransferase to a position in the centromeric region of mouse chromosome 4. *Cytogenet. Cell Genet.*, **54,** 172–174.

Shaper,N.L., Meurer,J.A., Joziasse,J.H., Chou,T.-D.D., Smith,E.J., Schnaar,R.L. and Shaper,J.H. (1997) The chicken genome contains two functional nonallelic β1,4-galactosyltransferase genes: chromosomal assignment to syntenic regions tracks fate of the two gene lineages in the human genome. *J. Biol. Chem.*, **272,** 31389–31399.

Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278,** 631–37.

Turkington,R.W., Brew,K., Vanaman,T.C. and Hill,R.L. (1968) The hormonal control of lactose synthetase in the developing mouse mammary gland. *J. Biol. Chem.*, **243,** 3382–3387.

Ulrich,J.T., Schenck,J.R., Rittenhouse,H.G., Shaper,N.L. and Shaper,J.H. (1986) Monoclonal antibodies to bovine UDP-galactosyltransferase. Characterization, cross-reactivity, and utilization as structural probes. *J. Biol. Chem.*, **261,** 7975–7981.

Yadav,S.P. and Brew,K. (1990) Identification of a region of UDP-galactose:N-acetylglucosamine β4-galactosyltransferase involved in UDP-galactose binding by differential labeling. *J. Biol. Chem.*, **265,** 14163–14169.

Yadav,S.P. and Brew,K. (1991) Structure and function in galactosyltransferase: sequence locations of α-lactalbumin binding site, thiol groups, and disulfide bond. *J. Biol. Chem.*, **266,** 698–703.