THE JOURNAL OF BIOLOGICAL CHEMISTRY Vol. 272, No. 50, Issue of December 12, pp. 31389–31399, 1997
*Printed in U.S.A.*

# 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*

(Received for publication, July 21, 1997)

**Nancy L. Shaper‡, Janet A. Meurer‡§¶, David H. Joziasse‖, T-D. D. Chou‡, Eugene J. Smith**, Ronald L. Schnaar§, and Joel H. Shaper‡§‡‡**

*From the ‡Cell Structure and Function Laboratory, The Oncology Center and §Department of Pharmacology and Molecular Sciences, School of Medicine, The Johns Hopkins University, Baltimore, Maryland 21287-8937, the ‖Department of Medical Chemistry, Vrije Universiteit, NL-1081-BT Amsterdam, The Netherlands, and the **United States Department of Agriculture, Agricultural Research Service, East Lansing, Michigan 48823*

**Two distinct but related groups of cDNA clones, CKβ4GT-I and CKβ4GT-II, have been isolated by screening a chicken hepatoma cDNA library with a bovine β1,4-galactosyltransferase (β4GT) cDNA clone. CKβ4GT-I is predicted to encode a type II transmembrane glycoprotein of 41 kDa with one consensus site for *N*-linked glycosylation. CKβ4GT-II is predicted to encode a type II transmembrane glycoprotein of 43 kDa with five potential *N*-linked glycosylation sites. At the amino acid level, the coding regions of CKβ4GT-I and CKβ4GT-II are 52% identical to each other and 62 and 49% identical, respectively, to bovine β4GT. Despite this divergence in amino acid sequence, high levels of expression of each cDNA in *Trichoplusia ni* insect cells demonstrate that both CKβ4GT-I and CKβ4GT-II encode an α-lactalbumin-responsive, UDP-galactose:*N*-acetylglucosamine β4-galactosyltransferase.**

**An analysis of CKβ4GT-I and CKβ4GT-II genomic clones established that the intron positions within the coding region are conserved when compared with each other, and these positions are identical to the mouse and human β4GT genes. Thus CKβ4GT-I and CKβ4GT-II are the result of the duplication of an ancestral gene and subsequent divergence. CKβ4GT-I maps to chicken chromosome Z in a region of conserved synteny with the centromeric region of mouse chromosome 4 and human chromosome 9p, where β4-galactosyltransferase (EC 2.4.1.38) had previously been mapped. Consequently, during the evolution of mammals, it is the CKβ4GT-I gene lineage that has been recruited for the biosynthesis of lactose. CKβ4GT-II maps to a region of chicken chromosome 8 that exhibits conserved synteny with human chromosome 1p. An inspection of the current human gene map of expressed sequence tags reveals that there is a gene noted to be highly similar to β4GT located in this syntenic region on human chromosome 1p. Because both the CKβ4GT-I and CKβ4GT-II gene lineages are detectable in mammals, duplication of the ancestral β4-galactosyltransferase gene occurred over 250 million years ago in an ancestral species common to both mammals and birds.**

β1,4-Galactosyltransferase (β4GT; EC 2.4.1.38)[1] 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 (1). In mammals, β4GT 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 β4GT and the noncatalytic protein α-lactalbumin, a mammalian protein expressed exclusively in lactating mammary epithelial cells (2–4). Interestingly, β4GT from nonmammalian species such as chicken (4) and plant (5) can also functionally interact with α-lactalbumin *in vitro*, indicating that the α-lactalbumin binding domain in β4GT predates the rise of mammals.

We have reported that the murine β4GT gene is unusual in that it specifies two size sets of mRNAs in somatic cells of ~3.9 and ~4.1 kb. These two transcripts arise as a consequence of initiation at two different sets of start sites that are separated in the first exon by ~200 bp. Because the respective start sites are positioned either upstream of the first of two in-frame ATGs (4.1 kb) or between these two in-frame ATGs (3.9 kb), translation of each mRNA results in the synthesis of two structurally related, *trans*-Golgi-resident protein isoforms that differ only in the length of their $NH_2$-terminal cytoplasmic domain (6). The identical structural features are also found in the bovine (7) and human β4GT gene (8, 9), suggesting that they

* This work was supported in part by National Institutes of Health Grants GM38310 and CA45799 (to J. H. S.), Grant HD14010 (to R. L. S.) and Human Frontiers Science Program RG-414/94M (to D. H. J.). A preliminary report of this work has been presented (14).The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

*The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EBI Data Bank with accession number(s) U19890 (CKβ4GT-I) and U19889 (CKβ4GT-II).*

¶ Supported in part by National Institutes of Health Training Grant CA 09243 in partial fulfillment of the degree of Doctor of Philosophy at The Johns Hopkins University. Present address: Berlex Biosciences, Richmond, CA 94804.

‡‡ To whom reprint requests should be addressed: The Johns Hopkins School of Medicine, Oncology Center, Room 1-127, 600 N. Wolfe St., Baltimore, MD 21287-8937. Fax: 410-550-5499; E-mail: jshaper@welchlink.welch.jhu.edu.

[1] The abbreviations used are: β4GT, β1,4-galactosyltransferase (EC 2.4.1.38); murine and human β4GT refer to the α-lactalbumin responsive, UDP-galactose:*N*-acetylglucosamine β4-galactosyltransferase that has been mapped to the centromeric region of mouse chromosome 4 and human chromosome 9p13, respectively; bovine β4GT is the corresponding bovine homologue of the mouse and human galactosyltransferase; bp, base pair(s); IREBP, iron response element-binding protein; JF, Jungle Fowl; kb, kilobase(s); nt, nucleotide; RPL5, ribosomal protein L5; WL, White Leghorn; PCR, polymerase chain reaction.

may be a distinguishing characteristic of all corresponding mammalian β4GT genes.

We have established that murine somatic tissues predominantly use the 4.1-kb transcriptional start site (10). The only exception to this general pattern is found in the mid- to late pregnant and lactating mammary gland, where the 3.9-kb transcriptional start site is preferentially used. This switch to the predominant use of the 3.9-kb start site is coincident with the cellular requirement for increased β4GT enzyme levels in preparation for lactose biosynthesis. These observations, combined with a detailed promoter analysis, experimentally 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 (10, 11). 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 β4GT for lactose biosynthesis (11). One prediction of this model is that, because β4GT in nonmammalian vertebrates functions exclusively in a housekeeping (glycan biosynthesis) role, the nonmammalian gene will exhibit only one (or one set of) clustered transcriptional start site(s), characteristic of many housekeeping genes. To test this prediction and to begin to generate a data base for comparing the amino acid sequence of β4GT from diverse species to identify essential amino acids for structure-function correlations, we have isolated and characterized the β4GT gene from a nonmammalian vertebrate, the chicken.

Based on the isolation of full-length cDNA and genomic clones and on the expression of enzymatically active recombinant protein, we report that the chicken genome contains two functional, divergent β4GT genes, termed CKβ4GT-I and CKβ4GT-II, and that each encodes an α-lactalbumin-responsive β4GT homologue. CKβ4GT-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 9p, where β4-galactosyltransferase had previously been mapped (12, 13). Consequently, it is the CKβ4GT-I ancestral lineage that has evolved into the mammalian β4GT gene that is recognized to function in lactose biosynthesis. In contrast, CKβ4GT-II maps to chicken chromosome 8, in a region that is syntenic with human chromosome 1p, where a set of expressed sequence tags, noted to be highly similar to β4GT, have been mapped. Because both the CKβ4GT-I and CKβ4GT-II gene lineages are detectable in mammals, duplication of the ancestral β4-galactosyltransferase gene occurred over 250 million years ago in an ancestral species common to both mammals and birds.

## MATERIALS AND METHODS

*Reagents*—Restriction enzymes, reverse transcriptase, T4 DNA ligase, the Klenow fragment of DNA polymerase I, S1 nuclease, and polynucleotide kinase were from Life Technologies, Inc. or New England Biolabs. *Taq* DNA polymerase was from Boehringer Mannheim. λgt10 arms and λ packaging extracts were from Promega Corp. *Eco*RI linkers were from Collaborative Research. Radioactive nucleotides were from Amersham Corp.

*Plasmids and Cell Lines*—The chicken MSB-1 and T-249 cell lines were obtained from Dr. W. Earnshaw (Institute of Cell and Molecular Biology, University of Edinburgh, Scotland), and grown in Dulbecco's modified Eagle's medium, high glucose supplemented with 10% fetal calf serum. The MSB-1 cell line is derived from a Marek's disease virus-induced lymphoma (15). The T-249 cell line was isolated from a liver tumor produced by the MC29 strain of avian leukosis virus (16). Culture media was supplemented with 100 units/ml penicillin and 50 μg/ml streptomycin. The cells were maintained at 37 °C in a 5% $CO_2$ atmosphere.

*Construction and Screening of a Chicken Hepatoma cDNA Library*—Double-stranded cDNA, prepared from 10 μg of T-249 poly(A)$^+$ RNA, was used to construct a λgt10 cDNA library using previously described procedures (6). Approximately $1 \times 10^6$ recombinant phage were screened using the bovine β4GT cDNA clone 7A as the probe. This clone contains ~1 kb of coding sequence plus ~300 bp of 3′-untranslated sequence (17).

*Construction of Recombinant Transfer Vector and Recombinant Baculovirus*—High Five cells (BTI-TN-5B1–4), derived from *Trichoplusia ni* egg cell homogenates were obtained from Invitrogen, and were grown in Hanks' TNM-FH medium (Sera-Lab Ltd, United Kingdom) supplemented with 10% fetal calf serum (Life Technologies). Linearized, wild-type *Autographa californica* virus DNA containing a deletion in an essential viral gene (BaculoGold DNA) was obtained from Pharmingen (San Diego, CA). The plasmid pVT-Bac was kindly donated by Dr. T. Vernet (Biotechnology Research Institute, Montreal, Canada). This transfer vector, which contains the signal peptide of the mellitin gene upstream of a multiple cloning site, was used for the construction of recombinant plasmids (18).

Plasmid pVT-Bac was digested with *Bam*HI and blunted using T4 DNA polymerase. A 980-bp *Ava*I fragment from CKβ4GT-I cDNA clone 33A, which contains the coding sequence from amino acid residue 45 and includes the stop codon, was blunted with T4 DNA polymerase and ligated into the vector. Plasmid pVT-Bac was digested with *Sma*I and *Eco*RI and ligated with a 2.2-kb *Acc*I/*Eco*RI fragment from CKβ4GT-II cDNA clone 25B that contains the coding sequence starting from amino acid residue 36 and includes the stop codon. Each clone was sequenced across the junction point.

Recombinant baculovirus was produced as described in the BaculoGold manual supplied by the manufacturer. Infection of *T. ni* insect cells with the recombinant virus results in the secretion of a soluble form of each chicken polypeptide. Cleavage by the signal peptidase results in one additional amino-terminal residue (Asp) in the polypeptide encoded by clone 33A and four additional residues (Asp-Pro-Ser-Pro) in the polypeptide encoded by clone 25B.

*Production of Recombinant Enzyme and Product Characterization—T. ni* cells were infected at a multiplicity of infection of 5 with recombinant baculovirus. At 72 h postinfection, the medium was collected and centrifuged to remove detached cells. Galactosyltransferase assays were carried out directly on aliquots (3–10 μl) of the medium in a final reaction volume of 50 μl containing 1.25 μmol of GlcNAc as the acceptor substrate, 5 μmol of Tris-maleate buffer, pH 6.8, 1 μmol of $MnCl_2$, 0.4 mg of Triton X-100, 200 nmol of ATP, 1 μmol of γ-galactono-1,4-lactone (Sigma), 25 μg of bovine serum albumin, and 25 nmol of UDP-[$^3$H]Gal (1.1 Ci/mol). The reaction mixtures were incubated at 37 °C for 15 min, and the product was isolated by ion exchange chromatography and quantified as described previously (19).

An aliquot of the radioactive product was subjected to high pH anion exchange chromatography with pulsed amperometric detection. The system consisted of a Dionex Bio-LC gradient pump, a Carbopac-100 column (4 × 250 mm), and a model PAD 2 detector. The following pulse potentials and durations were used for detection: $E_1 = 0.05$ V ($t_1 = 480$ ms); $E_2 = 0.60$ V ($t_2 = 120$ ms); $E_3 = 0.60$ V ($t_3 = 60$ ms). Samples were dissolved in 0.1 M NaOH, and the column was eluted isocratically with 0.1 M NaOH for 10 min, after which a gradient was applied that increased the concentration of sodium acetate in 0.1 M NaOH by 2.5 mM/min. The flow rate was 1 ml/min. The eluate was collected in 0.5-ml fractions, and radioactivity was counted in the individual fractions. The elution position of the radioactive product was compared with that of the reference compounds: Gal, Galβ1,6GlcNAc, Galβ1,3GlcNAc, and Galβ1,4Glc, as determined by pulsed amperometric detection.

The effect of α-lactalbumin on the acceptor preference of each recombinant enzyme was evaluated by performing the standard galactosyltransferase assay as described above, using either GlcNAc or Glc as the acceptor, in the presence of increasing α-lactalbumin concentrations (0.1–2.0 mg/ml).

*Isolation of Chicken Genomic Clones*—A λFIX chicken genomic library (kindly provided by Dr. C. B. Thompson, Howard Hughes Medical Institute, University of Chicago) and a chicken cosmid genomic library (Stratagene) were screened sequentially with the CKβ4GT-I and CKβ4GT-II cDNA clones, as described previously (20). The genomic inserts were characterized by restriction mapping and Southern blot analysis using exon-specific, $^{32}$P-labeled oligonucleotide probes. Specific genomic restriction fragments were subcloned and partially sequenced using exon-specific oligonucleotide primers. To subclone the 3.5-kb *Sst*I and 2.1-kb *Sst*I CKβ4GT-I genomic fragments containing exons 1 and 2, respectively, it was necessary to transform STBL-2 cells (Life Technologies), which are used to grow clones prone to deletion.

*Northern and Southern Blot Analysis*—RNA and DNA were isolated from T-249 cells by the guanidine isothiocyanate method of Chirgwin *et al.* (21). RNA was also isolated from MSB-1 cells and various tissues of a young female White Leghorn chicken. The isolation of poly(A)$^+$ RNA and the Northern and Southern blot analyses were carried out as described previously (17).

*S1 Nuclease Analysis*—S1 nuclease protection assays were performed as described previously (6). A 656-bp *Pvu*II-*Nar*I CKβ4GT-I genomic DNA fragment was isolated that flanked the anticipated transcriptional start site(s). The *Nar*I cleavage site corresponded to nt +227 in Fig. 1*A*. A single-stranded probe complementary to the CKβ4GT-I transcribed sequence was prepared by primer extension of an M13mp18 clone in the presence of [$^{32}$P]dATP and Klenow polymerase. The probe contained an additional 87 bp of polylinker sequence. A 305-bp *Not*I CKβ4GT-II genomic DNA fragment was isolated that spanned the 5′-end of the CKβ4GT-II sequence and a single-stranded probe, containing 97 bp of polylinker sequence, was prepared as described above. After purification, probe hybridization to MSB-1 and T-249 RNA was carried out at 62 °C overnight. The samples were digested with S1 nuclease, and the products were analyzed on a 7% polyacrylamide, 8 M urea gel.

*DNA Sequence and Computer Analyses*—Double-stranded, dideoxy-DNA sequencing was performed using the Sequenase kit from U.S. Biochemical Corp. The M13, T7, and T3 primers as well as synthetic oligonucleotides were used as sequencing primers. Oligonucleotides were synthesized by the Johns Hopkins Core Facility. Sequences were analyzed using MacVector from International Biotechnologies, Inc. or the GeneWorks Nucleic Acid and Protein Sequencing Analysis program from IntelliGenetics, Inc.[2]

*Oligonucleotide Primers*—The oligonucleotide primers used to determine the intron/exon boundaries for the CKβ4GT-I genomic clone were as follows: EX 1F, 5′-CCCGGACCGGTCTCGGCAC-3′; EX 2F, 5′-TACATGCACCCAATTCTTCA-3′; EX 2R, 5′-GACCTCAGGGTTTGTGCTCGC-3′; EX 3F, 5′-TACAAGTGCTACAGCCAAC-3′; EX 3R, 5′-AGCTTCCGTGAATCCTACA-3′; EX 4F, 5′-AAGATCAATGGGTTTCC-3′; EX 4R, 5′-TGCTCAAGGCAGACACACCTCC-3′; EX 5F, 5′-TGGGAAATGCAGAATGATT-3′; EX 5R, 5′-TCTGCATTTCCCAATGACAG-3′; EX 6R, 5′-AGCGAGTTCAAGCCATCAGA-3′.

The oligonucleotide primers used to determine the intron/exon boundaries for the CKβ4GT-II genomic clone were as follows: EX 0F, 5′-GCCGCGGGAGGAGGTGGC-3′; EX 1F, 5′-AACACCAACCGCTCCGTCAC-3′; EX 1R, 5′-AAGAGCAACCTGGTCAT-3′; EX 2F, 5′-CACTACCTGCACCCCATC-3′; EX 2R, 5′-TCAGGGTTCTCCCGTTGCAC-3′; EX 3F, 5′-AACCTGTACCGCTGCTATGA-3′; EX 3R, 5′-TACTCCTCGTCATCCTTCAGC-3′; EX 4F, 5′-AAGATCAATGGGTTTCC-3′; EX 4R, 5′-TGCTCAGCCCAGAGACACCTCC-3′; EX 5F, 5′-ATGGGGAGGTATCGCATGA-3′; EX 5R, 5′-TCATGCGATACCTCCCCAT-3′; EX 6R, 5′-TGATCCCATCCCGCTTCAT-3′.

*Localization of CKβ4GT-I and CKβ4GT-II in the Chicken Genome*—Each CKβ4GT gene was mapped using a mismatched primer PCR approach based on nucleotide substitutions found in either the Jungle Fowl (JF) or White Leghorn (WL) product (23). To map the β4GT genes, DNA from 52 progeny of the East Lansing reference population ((JF × WL) × WL) were used to follow segregation of the JF allele. When base substitutions were found, 3′-mismatched primers were designed to preferentially amplify only the JF allele. Because the WL is recurrent in this back-cross, only segregation of the JF can be scored.

PCR primers that initially amplified a region in the second intron of CKβ4GT-I were 5′-CAGGTGAGGGGTGCTGAGA-3′ (forward) and 5′-AGGCAGTCGTGAAAGAGA-3′ (reverse). The 530-bp product was sequenced, and an A-G transition was found between JF and WL. A JF allele-specific 3′-mismatched reverse primer, together with the original forward primer, amplified a 210-bp product; the parental WL allele was not amplified. PCR primers that initially amplified a region in the 3′-untranslated region of CKβ4GT-II were 5′-CAGACAGAGGGAGGGGAC-3′ (forward) and 5′-AGGGACACGCACACAGCA-3′ (reverse). The 410-bp PCR product was sequenced, and an A-G transition was found between JF and WL. A JF allele-specific 3′-mismatched reverse primer, together with the original forward primer was used to amplify a 257-bp product; the parental WL allele was not amplified.

## RESULTS

*Isolation and Characterization of Two Chicken Homologues of β4GT*—The strategy we used to clone chicken β4GT was to construct a λgt10 nonexpression cDNA library using poly(A)$^+$ RNA isolated from the T-249 cell line and to screen it with our bovine β4GT cDNA probe. Our choice of T-249 cells for library construction and the bovine probe for library screening was based on two considerations. First, by direct enzymatic assay, T-249 cells exhibited a 3-fold higher level of β4GT activity compared with MSB-1 cells. Second, Northern blot analysis of T-249 poly(A)$^+$ RNA revealed a broad hybridization positive band of ~2.5 kb using the bovine β4GT probe. This latter result indicated that there was sufficient similarity in the nucleotide sequence to permit direct screening with the bovine probe.

Approximately $1 \times 10^6$ independent recombinants were subsequently screened, resulting in the isolation of 18 cDNA clones. The six largest inserts were subcloned and partially sequenced. This preliminary analysis, in combination with partial restriction endonuclease mapping of the other 12 isolates, revealed the presence of two distinct groups of clones, CKβ4GT-I (nine clones) and CKβ4GT-II (nine clones).

*Nucleotide Sequence and Translated Amino Acid Sequence of CKβ4GT-I*—The complete nucleotide sequence of clone 33A is shown in Fig. 1*A*. The first in-frame ATG encoding a long open reading frame (located at nt +1 to +3) was present in a sequence context appropriate for translation initiation (23) and therefore was designated the initiating Met. The entire 3′-untranslated region (1137 bp) is present in this clone, since a consensus polyadenylation signal (ATTAAA) is present 15–20 bp upstream of a 21-nt poly(A) tail.

As discussed below, the additional 5′-untranslated sequence (−17 to −210), shown in Fig. 1*A*, was subsequently obtained after carrying out S1 analysis on a fragment derived from an appropriate genomic clone.

The coding region of clone 33A is 66% identical at the nucleotide level, and 62% at the amino acid level, to the corresponding bovine sequence. Translation predicts a type II, membrane-bound, potentially glycosylated protein of 362 amino acids with an $NH_2$-terminal cytoplasmic domain of 16 amino acids, a single transmembrane domain of 20 amino acids (assuming that the Gly residue at position 36 defines the COOH-terminal boundary of this domain), a stem region of ~55 amino acids, and a COOH-terminal domain of 271 amino acids. One *N*-linked glycosylation consensus site is located at Asn$^{56}$. The length of this chicken β4GT homologue is 40 amino acids shorter than the predicted long protein isoform of bovine β4GT due to multiple small deletions in the stem region combined with a cytoplasmic domain that is eight amino acids shorter.

*Nucleotide Sequence and Translated Amino Acid Sequence of CKβ4GT-II*—The complete nucleotide sequence of CKβ4GT-II (clone 25B) is shown in Fig. 1*B*. The first in-frame ATG codon of the longest open reading frame (located at nt +1 to +3) was assigned as the initiating Met, based on Kozak's rules for translation initiation (24) and the fact that an upstream in-frame termination codon (TGA) is present at nt −117 to −115. Consequently, this cDNA clone contains 209 bp of 5′-untranslated sequence plus a coding sequence of 1119 bp. The complete 3′-untranslated sequence (1100 bp) is also present in this clone, since a consensus polyadenylation signal (AATAAA) was located 17–22 bp upstream of a 65-nt poly(A) tail.

The coding region is 59% identical at the nucleotide level, and 49% identical at the amino acid level, to the corresponding bovine sequence. Translation predicts a type II, membrane-bound, potentially glycosylated protein of 373 amino acids with an $NH_2$-terminal cytoplasmic domain of 15 amino acids, a single transmembrane domain of 18 amino acids, a stem region of ~64 amino acids, and a COOH-terminal catalytic domain of 276 amino acids. Five *N*-linked glycosylation consensus sites are located at Asn$^{50}$ Asn$^{59}$, Asn$^{64}$, Asn$^{87}$, and Asn$^{358}$. The

[2] The UniGene data base is available through a site on the World Wide Web at http://www.ncbi.nlm.nih.gov/ (22).

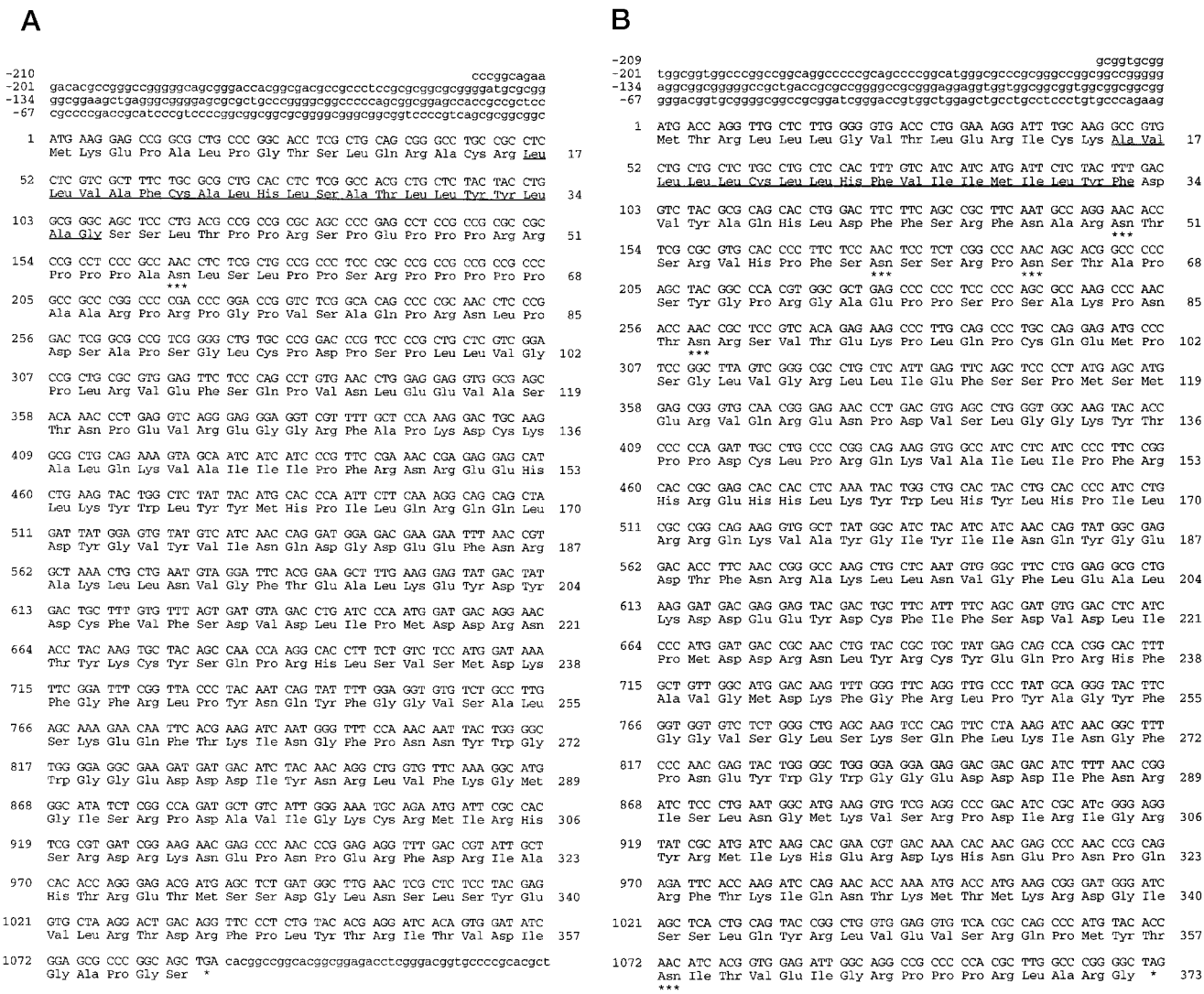


FIG. 1. **Nucleotide and predicted amino acid sequence of CKβ4GT-I and CKβ4GT-II.** The nucleotide sequence is numbered on the *left* with sequence upstream of the first in-frame ATG assigned negative numbers. The amino acid sequence is numbered at the *right*. The 5′- and 3′-untranslated sequences are in *lowercase letters*, while the nucleotide sequence corresponding to the coding sequence is in *capital letters*. The sequence encoding the putative transmembrane domain is *underlined*, and the predicted *N*-linked glycosylation sites are marked with a *triple asterisk* below the Asn residues. *A*, the full-length sequence of CKβ4GT-I (clone 33A) is shown, which begins at nt −16. The sequence from −17 to −210 was obtained after S1 analysis of an overlapping genomic clone. The polyadenylation sequence is *underlined* and in *boldface type*. *B*, the sequence of CKβ4GT-II (clone 25B) is shown with the polyadenylation sequence *underlined* and in *boldface type*.

length of this chicken β4GT homologue is 29 amino acids shorter than the predicted long protein isoform of bovine β4GT due to multiple deletions in the stem region, a cytoplasmic domain that is shorter by nine amino acids, and a six-amino acid extension at the COOH terminus.

At the nucleotide level, the coding region of CKβ4GT-II (clone 25B) is 61% identical to CKβ4GT-I (clone 33A). The respective 5′-untranslated regions that have a GC-content of ~84% were 42% identical; however, a number of gaps were required to obtain maximal alignment. In contrast, there is essentially no sequence identity between the respective 3′-untranslated regions.

*Northern Blot Analysis*—Northern blot analysis was carried out using RNA isolated from the T-249 cell line, used for cDNA library construction, to determine the size and number of transcripts corresponding to each cDNA clone. Since the nucleotide sequence of the coding region of CKβ4GT-I and CKβ4GT-II is 61% identical, the 3′-untranslated region of each transcript was used to probe the Northern blot. As indicated above, these regions share no sequence identity.

The CKβ4GT-I probe identifies an mRNA species of ~2.5 kb (Fig. 2*A*, *lane 1*). The minor band at ~4.3 kb is due to nonspe-

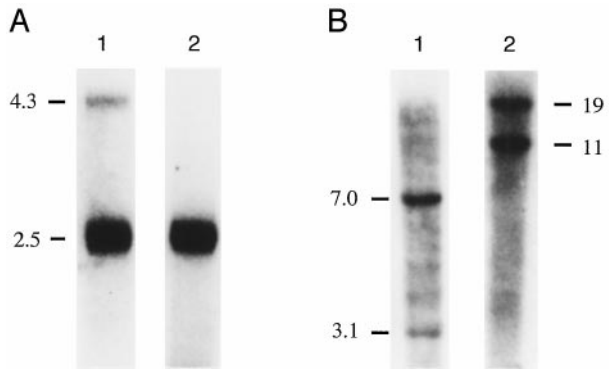


FIG. 2. **Northern and Southern blot analysis.** *A*, poly(A)$^+$ RNA (5 μg), isolated from T-249 cells, was resolved by electrophoresis on a formaldehyde-agarose gel, transferred to Nytran, and probed with a $^{32}$P-labeled 830-bp *Sma*I-*Eco*RI fragment derived from the 3′-untranslated region of CKβ4GT-I (*lane 1*). The minor hybridizing band at 4.3 kb is due to a low level of 28 S RNA remaining in the poly(A)$^+$ RNA fraction. After probe removal, the blot was rehybridized with a $^{32}$P-labeled 945-bp *Pst*I fragment derived from the 3′-untranslated region of CKβ4GT-II (*lane 2*). *B*, genomic DNA (15 μg) isolated from T-249 cells was digested with *Bam*HI, electrophoresed on an 0.8% agarose gel, transferred to Nytran, and hybridized with either the $^{32}$P-labeled 2.3-kb insert from clone 3A encoding CKβ4GT-I (*lane 1*) or the *Sst*II-*Eco*RI 2.3-kb fragment encoding CKβ4GT-II (*lane 2*).

cific hybridization of the probe to 28 S ribosomal RNA. Since CKβ4GT-I clone 33A (~2.2 kb) contains a consensus polyadenylation site and a poly(A) tail, the missing sequence (estimated to be ~150 bp) is from the 5′-end.

The CKβ4GT-I probe was subsequently removed from the Northern blot, which was then hybridized with the 3′-untranslated region of the CKβ4GT-II clone. A single transcript of ~2.5 kb was detected, suggesting that clone 25B represents the full-length transcript (Fig. 2*A*, *lane 2*).

*CKβ4GT-I and CKβ4GT-II Each Encode an Enzymatically Active, α-Lactalbumin-responsive β4-Galactosyltransferase*—To determine if each chicken cDNA encodes a β1,4-galactosyltransferase, constructs were assembled that fused the lumenal domain of either CKβ4GT-I or CKβ4GT-II to the signal sequence of mellitin. Expression of each cDNA in *T. ni* insect cells resulted in the secretion of enzymatically active, soluble enzyme. As shown in Table I, both CKβ4GT-I and CKβ4GT-II showed a relatively high galactosyltransferase activity using UDP-Gal as the donor and GlcNAc as the acceptor substrate. Furthermore, each recombinant galactosyltransferase is able to interact productively with α-lactalbumin as evidenced by the production of lactose. When UDP-Gal was replaced by equal concentrations of UDP-GalNAc, UDP-GlcNAc, or UDP-Glc, the activity was reduced to less than 1% of that measured with UDP-Gal. This low level of residual activity was comparable with that observed with affinity-purified bovine β1,4-galactosyltransferase (data not shown).

Since both β1,3- and β1,4-galactosyltransferases have been detected in chicken, the product formed using GlcNAc as the acceptor was characterized. On high pH anion exchange chromatography, the radioactive product migrated as a single peak, whose elution position corresponded to authentic Galβ1,4GlcNAc (retention time 7.2 min). No radioactivity was detected at the elution position of Galβ1,3GlcNAc (retention time 10.5 min). Collectively, these results establish that both CKβ4GT-I and CKβ4GT-II encode an α-lactalbumin-responsive UDP-Gal:GlcNAc-R β1,4-galactosyltransferase.

*Comparison of the Amino Acid Sequences of CKβ4GT-I, CKβ4GT-II, and Bovine β4GT*—The protein domain structure established for the cloned mammalian β4-galactosyltransferases consists of (i) a short $NH_2$-terminal cytoplasmic domain

TABLE I
*Expression of enzymatically active recombinant CKβ4GT-I and CKβ4GT-II in T. ni insect cells*

Recombinant CKβ4GT-I and CKβ4GT-II proteins were produced as secreted, soluble forms from *T. ni* insect cells and assayed for enzymatic activity as detailed under "Materials and Methods." Based on substrate preference and product characterization, both the CKβ4GT-I and CKβ4GT-II genes encode an α-lactalbumin-responsive, UDP-galactose: *N*-acetylglucosamine β4-galactosyltransferase. One unit of enzyme activity is defined as the amount of enzyme that catalyzes the transfer of 1 μmol of Gal from UDP-Gal to GlcNAc/min at 37 °C. The indicated percentages are the relative activities for the indicated assay, based on the results obtained with GlcNAc as the acceptor in the absence (−) of α-lactalbumin (α-LA), whose activity was set at 100%.

| | Transfer to GlcNAc | Transfer to Glu |
|---|---|---|
| | *milliunits/ml* | |
| CKβ4GT-I | | |
| α-LA (−) | 153.0 (100%) | 3.9 (2%) |
| α-LA (+) | 25.0 (16%) | 132.0 (86%) |
| CKβ4GT-II | | |
| α-LA (−) | 28.0 (100%) | 1.9 (6%) |
| α-LA (+) | 3.4 (12%) | 36.4 (130%) |

of 11 or 24 amino acids depending on the protein isoform (6) and (ii) a large COOH-terminal lumenal domain (~224 amino acids) containing the catalytic center, linked to a single transmembrane domain (20 amino acids) through a potentially glycosylated peptide segment of ~85 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 $Cys^{134}$ and $Cys^{247}$ (Ref. 25; see bovine sequence in Fig. 3, *solid arrowheads*). This loop plus adjacent sequence in the stem region (the stem region is defined as the amino acid sequence between the transmembrane domain and $Cys^{134}$) is involved in α-lactalbumin binding as established by protection studies (26) and antibody blocking studies (27, 28). (ii) The COOH-terminal 155-amino acid segment contains two polypeptides, in the vicinity of $Cys^{342}$ (Fig. 3, bovine sequence), that can be affinity-labeled with UDP-Gal analogues (26, 29) or have been implicated in substrate binding by site-directed mutagenesis (29).

In the context of this domain structure, a comparison of the amino acid sequence between each chicken β4GT homologue and a mammalian (bovine) β4GT is interesting (Fig. 3). The amino acid sequence of CKβ4GT-I and CKβ4GT-II is 62 and 49% identical, respectively, to the bovine β4GT sequence and only 52% identical to each other. Thus, CKβ4GT-I and CKβ4GT-II are as divergent from each other as they are from their mammalian counterpart. When the amino acid sequences of all three proteins are compared, the sequence identity is reduced to about 42%. The structural domains that are least conserved are the stem domain and the $NH_2$-terminal region of the cytoplasmic domain (Fig. 3). Of particular note, six of the seven Cys residues including the two Cys residues involved in intramolecular disulfide bond formation ($Cys^{134}$ and $Cys^{247}$ in the bovine sequence, Fig. 3) are conserved in the CKβ4GT-I and CKβ4GT-II sequences. The remaining Cys residue (Fig. 3, *open arrowhead*) is conserved only in CKβ4GT-I; in CKβ4GT-II, this Cys residue is replaced by Tyr. As discussed below, this fortuitous Cys to Tyr replacement is a useful marker to follow the evolutionary gene lineage of CKβ4GT-I and CKβ4GT-II in the human and mouse genomes.

*Southern Blot Analysis and Isolation of CKβ4GT-I and CKβ4GT-II Genomic Clones*—The comparison of the nucleotide and amino acid sequence suggested to us that each chicken homologue is encoded by a separate nonallelic gene that arose as a consequence of duplication of an ancestral gene followed by

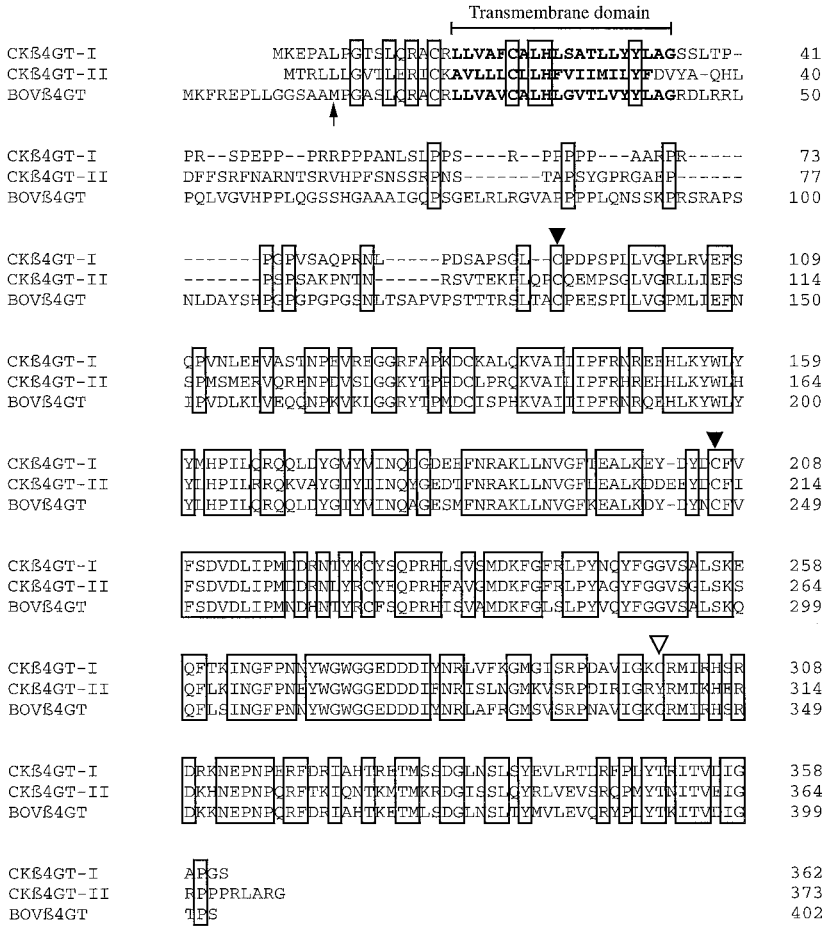


FIG. 3. **Alignment of the amino acid sequence of CKβ4GT-I, CKβ4GT-II, and bovine β4GT (*BOVβ4GT*).** The predicted amino acid sequences of CKβ4GT-I, CKβ4GT-II, and bovine β4GT are shown. *Boxed regions* indicate identical amino acids; *dashes* indicate introduced gaps. The putative transmembrane domain of CKβ4GT-I (residues 17–36), CKβ4GT-II (residues 16–33), and bovine β4GT (residues 25–44) is in *boldface type* and *overlined*. The *solid arrow* denotes the second in-frame $Met^{14}$ in bovine β4GT. The *solid arrowheads* indicate $Cys^{134}$ and $Cys^{247}$ (bovine sequence), which form the only intramolecular disulfide bond present in the molecule (25). The *open arrowhead* indicates $Cys^{342}$ (bovine sequence) that is replaced by a Tyr in CKβ4GT-II. Note that the second in-frame Met ($Met^{14}$ in the bovine sequence), which serves as the translational start site for the short β4GT protein isoform, is not present in either chicken sequence.

divergence. If this is the case, two predictions can be made. First, Southern analysis should result in two distinct patterns of restriction fragments. Second, the intron/exon boundaries within the coding region of each homologue should be identical. To test the first prediction, a Southern blot containing *Bam*HI-digested T-249 genomic DNA was hybridized with CKβ4GT-I or CKβ4GT-II. As seen in Fig. 2*B*, CKβ4GT-I hybridizes to 7- and 3.1-kb bands. In contrast, CKβ4GT-II hybridizes to 14- and 10-kb bands. This dissimilar pattern confirms that the chicken genome contains separate genes encoding CKβ4GT-I and CKβ4GT-II.

To confirm the second prediction, chicken genomic clones representing each homologue were isolated and characterized to determine intron/exon boundaries. A λFIX amplified chicken genomic library was screened using the CKβ4GT-I cDNA probe; 13 clones (~13 kb), which proved to be identical, were isolated. Initial characterization including restriction map analysis, in conjunction with Southern analysis, with probes from different regions of the coding sequence, established that the genomic clone (clone 1a), contained sequence limited to exons 3–6 (Fig. 4*A*). Consequently, a cosmid library was subsequently screened and two overlapping genomic clones (clones 6b and 14e) of ~40 kb were isolated that contained the missing upstream sequence.

The λFIX chicken genomic library was also screened with the CKβ4GT-II cDNA probe, and eight overlapping clones were identified. The longest clone (clone 2a; 14 kb) contained the entire coding and 3′-untranslated sequence, but the majority of the 5′-untranslated sequence was missing (Fig. 4*B*). The other seven clones also lacked this sequence. Subsequent screening of the cosmid library resulted in the isolation of clone 5a (~40 kb), which contained the missing sequence.

To establish the intron/exon boundaries, subcloned fragments of the genomic clones were sequenced using exon-specific oligonucleotide primers. Since intron/exon boundaries within the protein coding region are generally conserved across species lines, the CKβ4GT-I and CKβ4GT-II exon-specific primers used for sequencing were chosen based on the intron/exon boundaries established for the murine (30) and human β4GT gene (9). The sequence at the intron/exon boundaries determined for CKβ4GT-I and CKβ4GT-II along with the corresponding murine sequence is shown in Table II. Based on this analysis, it is clear that the two chicken β4GT genes share identical intron/exon boundaries with each other and their mammalian homologues, an observation that supports the notion of a gene duplication of an ancestral gene. However, unlike the mammalian β4GT gene, in which the entire 5′-untranslated region and first 415 bp of coding sequence are present on exon 1, the CKβ4GT-II gene has one intron within its 5′-untranslated region, positioned at nt −45.

*Expression of CKβ4GT-I and CKβ4GT-II in Various Adult Chicken Tissues*—The presence of two functional β4GT genes in the chicken would suggest that each is regulated in a tissue-specific manner. To examine this possibility, a Northern blot containing poly(A)$^+$ RNA isolated from brain, kidney, liver, lung, spleen, and pancreas of a female adult chicken was prepared and hybridized sequentially, with a probe derived from the 3′-untranslated region of each clone. As seen in Fig. 5, the steady state levels of the CKβ4GT-II mRNA are significantly higher in the panel of somatic tissues examined. Somewhat surprisingly, transcript levels are also high in the brain; this is in contrast to mice and humans, where β4GT mRNA levels are about 10-fold lower in the brain as compared with other somatic tissues.

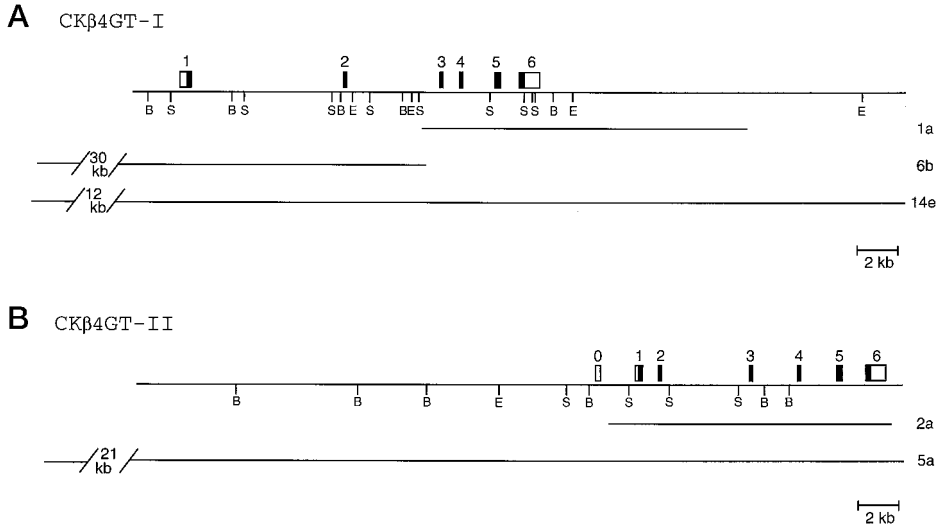


FIG. 4. **Genomic organization and partial restriction map of the CKβ4GT-I and the CKβ4GT-II gene.** The CKβ4GT-I and CKβ4GT-II genes are distributed over 20 and 16 kb of genomic DNA, respectively. In contrast, the murine (30) and human (9) β4GT genes span ~50 kb of genomic DNA. Exons, indicated by the *boxes*, are numbered 1–6 in agreement with the convention established for the mammalian β4GT genes (30). Exon 0 represents the additional 5′-exon present in CKβ4GT-II. The *solid boxes* and the *open boxes* correspond to the protein coding sequence and the untranslated sequence, respectively. *Bam*HI (*B*), *Eco*RI (*E*), and *Sst*I (*S*) sites are shown. The relative position of each exon within an indicated *Sst*I fragment is approximate.

*Analysis of the Transcriptional Start Site*—Using the rapid amplification of cDNA ends procedure, attempts to obtain the ~150 bp of sequence estimated to be missing from the 5′-untranslated region of CKβ4GT-I clone 33A were unsuccessful, probably due to the high GC content (84%) of this region. Therefore, S1 analysis was used as an alternative strategy to obtain this sequence and simultaneously map the transcriptional start site(s). This strategy was based in part on the fact that for the mammalian β4GT, the complete 5′-untranslated region (~200 bp) and initial 415 bp of coding sequence are located on the first exon; consequently, we felt that it was reasonable to assume that the organization of the 5′-end of the CKβ4GT-I gene would be similar.

A single-stranded 656-nt probe, containing genomic sequence that spans the 5′-end of CKβ4GT-I clone 33A, was hybridized to RNA isolated from MSB-1 and T-249 cells. A protected product of ~470 bp was observed (data not shown), showing that the 5′-untranslated region of CKβ4GT-I is ~210 nt in length and that this entire untranslated sequence is contiguous with the first 302 bp of coding sequence on exon one. The length of the protected product (~470 ± 20 bp) precludes a definitive conclusion as to the presence of a single start site or, alternatively, a set of clustered start sites within this short ~40-bp genomic sequence.

When a single-stranded 402-nt probe containing genomic sequence that spanned the 5′-end of CKβ4GT-II was hybridized to MSB-1 and T-249 RNA, protected products of 165, 175, 190, and 220 bp were observed (data not shown). This result shows that the CKβ4GT-II gene contains a set of closely spaced clustered start sites spanning ~55 bp. Use of each of these start sites would yield transcripts with a 5′-untranslated region of 210, 220, 235, and 255 nt, respectively; clone 25B represents the use of the most proximal start site.

*Chromosomal Assignment Permits Tracking the CKβ4GT-I and CKβ4GT-II Ancestral Lineages in the Human Genome*—β4GT has been cloned from three different mammals: cows, mice, and humans. From these published reports, only a single group of cDNA clones with highly conserved coding sequences (>85%), indicative of a single gene, have been isolated. These observations, coupled with the identification and characterization of two functional α-lactalbumin-responsive β4GT genes in the chicken genome, which arose as a consequence of duplication of an ancestral gene, raise two questions. First, which chicken β4GT gene lineage is the source of the mammalian β4GT gene that is recognized to function in both glycan biosynthesis and lactose biosynthesis? Second, is there a second functional β4GT gene in the mammalian genome?

A comparison of the respective coding regions of the two chicken β4GT genes with their mammalian counterpart suggests an answer to the first question. As previously discussed (Fig. 3), the coding sequence of CKβ4GT-I, at the nucleotide level, exhibits a somewhat greater sequence identity (66%) to the mammalian (bovine) gene compared with CKβ4GT-II (59%). Based on this analysis, it would appear that the CKβ4GT-I lineage gave rise to the well characterized mammalian β4GT gene that was recruited for lactose biosynthesis.

However, a more definitive approach, based on chromosomal assignment, can be used to trace the fate of each chicken β4GT gene lineage in the mammalian (human) genome. This strategy takes advantage of the comparative gene maps that have been established between different species, revealing regions of conserved synteny. These regions define groups of genes that are located together in close proximity on a chromosome. Therefore, given the location of a gene in one species, the location in another can be predicted.

Human β4GT maps to chromosome 9p13 (13) and to the centromeric region of mouse chromosome 4 (12) in a region that shows conserved synteny with aconitase. Aconitase also has a second function, in that it acts as the iron response element-binding protein (IREBP) (31). In the chicken, aconitase I/IREBP has been mapped to chromosome Z (32). Using allele-specific primers, we found that CKβ4GT-I maps to chicken chromosome Z to a region within 2 centimorgans of aconitase I/IREBP (Fig. 6*A*). This assignment, then, unequivocally establishes that the CKβ4GT-I gene lineage gave rise to the previously characterized human and mouse β4GT gene.

We have also mapped the CKβ4GT-II gene to chicken chromosome 8. Ribosomal protein L5 (RPL5) also maps to this small chromosome (33). In the human genome, RPL5 maps to chromosome 1p31–32 (Fig. 6*B*).[3] We have been able to take advantage of the recently established human gene map of expressed sequence tags (UniGene data base (22) to determine if any sequence tags, with noted similarity to β4GT, are present on human chromosome 1p 31–32 near RPL5. In fact, a group of 10 sequence tags (*e.g.* accession numbers W07207 and AA453005), which delineate a partial mRNA of ~1.5 kb, have been mapped to this chromosomal region (Fig. 6*B*). This ~1.5-kb mRNA has an open reading frame that encodes a protein of 279 amino acids that corresponds to about 75% of the coding sequence (based on the CKβ4GT-II coding sequence), including the complete catalytic domain. At the nucleotide and amino acid level,

[3] N. Kenmochi, T. S. Kawaguchi, S. Rozen, E. Davis, N. Goodman, T. Tanaka, and D. C. Page, manuscript in preparation.

TABLE II

*The intron/exon boundaries in the CKβ4GT-I and CKβ4GT-II genes*

The intron/exon junctions of CKβ4GT-I and CKβ4GT-II are shown and are compared with those established for murine (MU) β4GT. Exon nucleotide sequence is represented by capital letters; intron sequence is represented by lowercase letters. The superscript number refers to the position of either the first or last nucleotide within a given exon.

| | | | |
|---|---|---|---|
| CKβ4GT-II | | Exon 0 | ---GCCGCG$^{-46}$gaaggc |
| CKβ4GT-I | | | ---TGCTCG$^{301}$gtgagt |
| MUβ4GT | | Exon 1 | ---TGCTCG gtaaga |
| CKβ4GT-II | cagcag$^{-45}$GATCGG--- | | ---GCTTAG$^{316}$gtgaga |
| CKβ4GT-I | ttccag$^{302}$TCGGAC--- | | ---AACCAG$^{537}$gtaaga |
| MUβ4GT | ctctag TTGGCC--- | Exon 2 | ---AATCAG gtgagg |
| CKβ4GT-II | ttgcag$^{317}$TCGGGC--- | | ---AACCAG$^{552}$gtgagt |
| CKβ4GT-I | aaccag$^{538}$GATGGA--- | | ---ATTTCG$^{725}$gtgaga |
| MUβ4GT | caccag GCTGGA--- | Exon 3 | ---GTTTAG gtaaga |
| CKβ4GT-II | cctcag$^{553}$TATGGC--- | | ---GTTCAG$^{743}$gtgggt |
| CKβ4GT-I | ttctag$^{726}$GTTACC--- | | ---CAACAG$^{848}$gtaaag |
| MUβ4GT | ctccag CCTGCC--- | Exon 4 | ---TAACAG gtaatg |
| CKβ4GT-II | ctgcag$^{744}$GTTGCC--- | | ---TAACCG$^{866}$gtagtg |
| CKβ4GT-I | cggcag$^{849}$GCTGGT--- | | ---GGAGAG$^{953}$gtgcgg |
| MUβ4GT | tggcag ATTAGT--- | Exon 5 | ---TCAGAG gtaatg |
| CKβ4GT-II | ctccag$^{867}$GATCTC--- | | ---GCAGAG$^{971}$gtgagc |
| CKβ4GT-I | ccgcag$^{954}$GTTTGA--- | | ---3′-end |
| MUβ4GT | tggtag GTTTGA--- | Exon 6 | ---3′-end |
| CKβ4GT-II | caccag$^{972}$ATTCAC--- | | ---3′-end |

this human β4GT homologue is 77 and 80% identical to CKβ4GT-II, respectively. From an inspection of the primary sequence, the features suggestive of an UDP-galactose:*N*-acetylglucosamine β4-galactosyltransferase are apparent. Specifically, the relative positions of the four cystinyl residues in the catalytic domain and the essential amino acids in the two polypeptides pointed out above that have been affinity-labeled with UDP-Gal analogues are present (26, 27). Last, at the position of the Cys$^{342}$ to Tyr substitution (Fig. 3, *open arrowhead*), which distinguishes the respective CKβ4GT-I and CKβ4GT-II gene lineages, a Tyr is present. The open question is whether this human CKβ4GT-II homologue on chromosome 1p still encodes an enzymatically active, α-lactalbumin-responsive β4-galactosyltransferase.

## DISCUSSION

The unanticipated result from this study was the demonstration that the chicken genome contains two functional, nonallelic β4GT genes (CKβ4GT-I and CKβ4GT-II), which encode distinct α-lactalbumin-responsive, enzymatically active proteins that are only 52% identical (Fig. 3). Based on the conservation of the intron-exon boundaries within the coding region among CKβ4GT-I, CKβ4GT-II, and the mammalian β4GT genes, it is clear that these chicken β4GT genes arose as a consequence of duplication of an ancestral gene and subsequent divergence. When did duplication of the ancestral "β4GT gene" occur, relative to the independent evolution of mammals and birds? In considering this question, it is essential to recall that current opinion holds that mammals and birds last shared a common ancestor ~250 million years ago.[4] Consequently, depending on the time of the gene duplication, relative to the divergence from their common ancestor, two different outcomes can be predicted. First, if duplication of the ancestral β4GT gene occurred after divergence, it would be anticipated that these two β4GT genes would be a distinguishing characteristic of the avian genome and would not be found in the mammalian genome. In contrast, if the gene duplication took place prior to the separation of mammals and birds from their common predecessor, then one would anticipate finding both the CKβ4GT-I and CKβ4GT-II gene lineages in the mammalian genome.

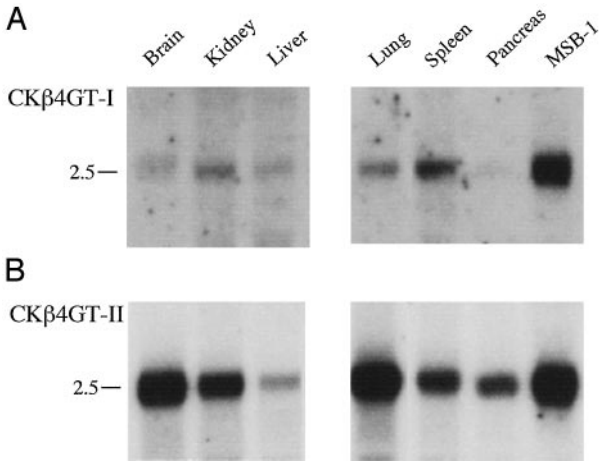


FIG. 5. **Expression of CKβ4GT-I and CKβ4GT-II in various adult chicken tissues.** Poly(A)$^+$ RNA (5 μg) was resolved on a formaldehyde agarose gel and transferred to Nytran. The blot was probed with an 830-bp $^{32}$P-labeled fragment derived from the 3′-untranslated region of CKβ4GT-I clone 33A (*panel A*). After probe removal, the blot was rehybridized with the 945-bp $^{32}$P-labeled fragment derived from the 3′-untranslated region of CKβ4GT-II clone 25B (*panel B*). Each probe was labeled to approximately the same specific activity; consequently, the steady state mRNA levels in the indicated tissues are directly comparable. RNA sizes in kb were determined relative to an RNA ladder.

As summarized in Fig. 6, we have mapped the CKβ4GT-I gene to chromosome Z, in a region that is syntenic with human chromosome 9p13, which is the chromosomal location of human β4GT. Consequently, it can be concluded that it was the CKβ4GT-I gene lineage that was recruited for lactose biosynthesis during the evolution of mammals. Additionally, CKβ4GT-II maps to chicken chromosome 8, in a region that is syntenic with human chromosome 1p, and where a set of expressed sequence tags with noted similarity to β4GT has recently been mapped. Thus, both the CKβ4GT-I and CKβ4GT-II lineages can be detected in the mammalian genome, indicating that duplication of the ancestral β4GT gene occurred at least 250 million years ago, prior to the divergence of mammals and avians from their common ancestor.

*The Human Genome Contains Additional CKβ4GT-II-related Genes*—Interestingly, four additional sets of expressed sequence tags with noted similarity to β4GT were also noted in the UniGene data base. Three sets map to human chromosomes 1q21-23, 3q13, and 18q11, respectively. The fourth set of expressed sequence tags has not yet been assigned a chromosomal position. From the available coding sequence, it is clear that three of these additional human β4GT-related genes encode a type II protein with a coding sequence that is ~40% identical with each other and with mouse or human β4GT.[2] (For the fourth set of sequence tags, mapped to 18q11, only the C-terminal 120 amino acids have been reported). An inspection of their primary sequence reveals that the relative positions of the six cystinyl residues are also conserved. Last, at the position of the Cys$^{342}$ to Tyr substitution (Fig. 3, *open arrowhead*), which distinguishes the respective CKβ4GT-I and CKβ4GT-II lineages, a diagnostic Tyr is present in each of the additional homologues. Consequently, it would appear that these four additional human β4GT homologues have arisen from multiple duplications within the CKβ4GT-II gene lineage.

Multiple mouse expressed sequence tags with noted similarity to β4GT have also been deposited in the dbEST data bank. Unfortunately, these mouse sequence tags have not been mapped; consequently, it is not possible to group these clones

[4] For a detailed overview of phylogeny, refer to The Tree of Life site on the World Wide Web at http://phylogeny.arizona.edu/tree/phylogeny.html.

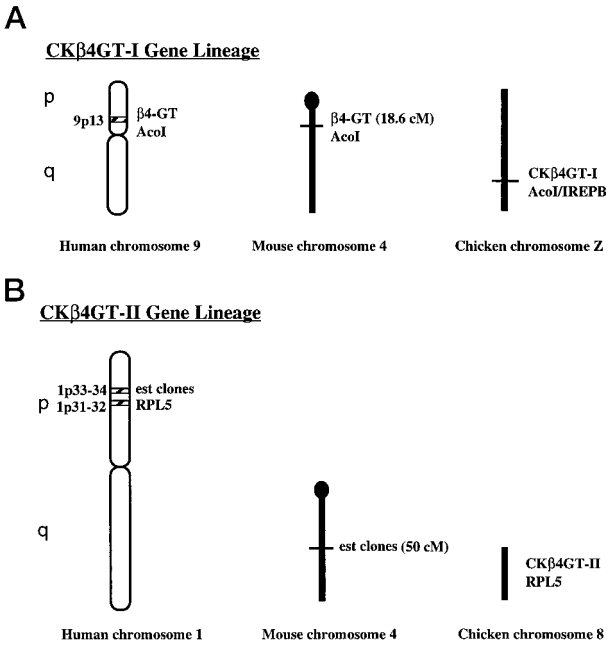


FIG. 6. **Chromosomal positions of the human and murine β4GT gene, the two chicken β4GT genes, and the corresponding synteny groups.** *A*, β4GT and aconitase I map to human chromosome 9p13 (*left*) and to mouse chromosome 4, ~18.6 centimorgans from the centromere (*middle*). The position of CKβ4GT-I and the aconitase I/IREBP gene is shown on the long arm of chicken chromosome Z (*right*). The Z chromosome of birds is the sex chromosome common to males (ZZ) and females (ZW). *B*, the position of RPL5 and the subset of expressed sequence tag (*est*) clones on human chromosome 1p is shown (*left*). The corresponding position in the mouse is predicted to be on chromosome 4, approximately 50 centimorgans from the centromere (*middle*). Both CKβ4GT-II and RPL5 have been mapped to the small chicken chromosome 8 (*right*).

by chromosomal assignment, as has been accomplished with the human sequence tags. The largest subset of these clones, based on essentially 100% sequence identity, represent the previously cloned β4GT gene that is located at the centromeric region of chromosome 4 (12). Among the remaining clones, there are a number of candidate subsets that potentially correspond to the human β4GT homologue located on chromosome 1p, which, based on conserved synteny, would be predicted to map to mouse chromosome 4, approximately 50 centimorgans from the centromere (Fig. 6*B*).

*Do the Human and Murine β4GT Homologues Encode β4GT Enzymatic Activity?*—Two laboratories have independently reported the generation of mice in which the β4GT gene has been inactivated by homologous recombination (34, 35). The null mice survive to term, and a significant percentage survive to maturity and are fertile. Interestingly, both groups have reported residual β4GT enzymatic activity in the range of ~5% in several tissues including the liver, testis, and brain, when GlcNAc (35) or asialylagalactotransferrin (34) was used as the acceptor sugar substrate. This residual β4GT enzymatic activity invites the conclusion that one or more of the new mouse β4GT homologues may in fact encode a β4-galactosyltransferase enzymatic activity. The enzymatic activity encoded by each of these human and mouse β4GT homologues is currently under investigation.

*The Golgi Retention Signal: Comparison of the Transmembrane Domain and Flanking Sequences between Chicken and Mammalian β4GT*—The current view is that the transmembrane domain plays a major role in the Golgi retention of type II membrane-bound proteins (reviewed in Refs. 36–38). For specific Golgi-resident glycosyltransferases, the sequences flanking the transmembrane domain may also be required for a fully functional retention signal. Interestingly, a comparison of the amino acid sequence of the $NH_2$-terminal regions of a group of resident proteins with a similar Golgi distribution has failed to reveal a sequence motif in common that could function as a Golgi retention signal.

An alternative approach to identifying essential amino acids within a functional domain is to compare the primary structure of the same protein from evolutionarily distant species. Using this strategy, one can establish the "mutations" allowed by nature consistent with maintenance of the functional domain. This approach is particularly applicable for an interspecies comparison of β4GT because the $NH_2$-terminal region including the lumenal stem domain (amino acids 1–92), in contrast to the COOH-terminal catalytic domain, exhibits the greatest divergence in primary structure (Fig. 3). Interestingly, within this region of divergence, the transmembrane domain and the nine amino acids of the cytoplasmic domain that immediately flank the transmembrane domain, stand out as being highly conserved. This point is further amplified by an inspection of the sequence alignment of the $NH_2$-terminal regions of the human, bovine, murine, CKβ4GT-I, and CKβ4GT-II β4GT polypeptides (Fig. 7). In the $NH_2$-terminal flanking sequence, four amino acids are identical and three are conservative replacements. Within the transmembrane domain, four amino acids are identical and six are conservative replacements. In contrast, the amino acids in the remainder of the cytoplasmic domain and the lumenal sequence flanking the transmembrane domain are not conserved. The fact that the indicated subset of amino acids distributed within the transmembrane and cytoplasmic domain have remained conserved over ~250 million years of evolution suggests that they may serve as a "functional unit" for retention of β4GT in the *trans*-Golgi.

*Absence of the 13-Amino Acid $NH_2$-terminal Extension in the Chicken β4GT Homologues*—Transcription of the murine β4GT gene in somatic tissues takes place at one of two different start sites that are separated by ~200 bp (Fig. 8). Use of these two transcriptional start sites results in a 4.1- and a 3.9-kb mRNA. The main difference between these two mRNAs is the length and extent of the predicted secondary structure of the respective 5′-untranslated regions (10). The 4.1-kb start site is positioned upstream of the first two in-frame ATGs, whereas the 3.9-kb start site is located between these two in-frame ATGs (Fig. 8). Consequently, translation of the 4.1- and 3.9-kb mRNAs results in the biosynthesis of two protein isoforms that differ only in the length of their respective $NH_2$-terminal cytoplasmic domains. The "long" and "short" β4GT protein isoforms have $NH_2$-terminal cytoplasmic domains of 24 and 11 amino acids, respectively.

The functional significance of this additional 13 amino acids has been the subject of much interest and speculation. Based on the conclusions from a number of investigators who showed that the transmembrane domain of β4GT is sufficient to retain a reporter protein in the Golgi compartment (reviewed in Refs. 36–38) and our demonstration that both β4GT protein isoforms were localized in the *trans*-Golgi compartment as assessed by immunoelectron microscopy, we have concluded that both isoforms are functionally equivalent Golgi-resident proteins (39). A contrasting viewpoint has been put forth by Shur and colleagues (40), who suggested that the 13-amino acid extension serves a functional role by overriding the *trans*-Golgi retention signal, thereby directing a small percentage or "portion" of this isoform to the cell surface. It was posited that, at the cell surface, the long β4GT isoform functions as a cell adhesion molecule by virtue of its ability to interact with the cytoskeleton via this 13-amino acid extension (41).

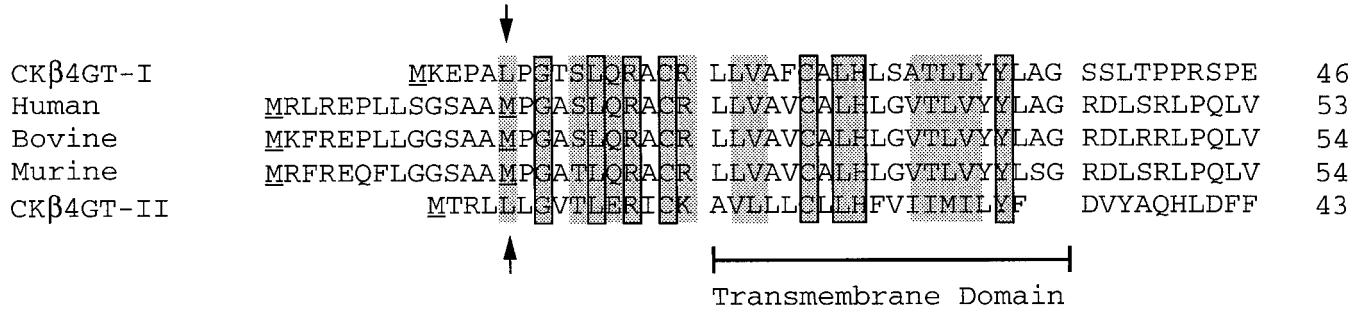


FIG. 7. **Alignment of the cytoplasmic, transmembrane domain, and lumenal flanking sequence of the mammalian and chicken β4-galactosyltransferase homologues.** Identical amino acids are *boxed*; similar amino acids are *shaded*. The initiating Met is *underlined*. The *arrows* indicate the position of the second in-frame Met in the mammalian β4GTs. The corresponding amino acid in CKβ4GT-I is a Leu (CTG), whereas in CKβ4GT-II it is a Leu (CTC). The amino acid residues in the respective transmembrane domains are indicated by the *horizontal line*. Note the subset of conserved amino acids distributed within the transmembrane and cytoplasmic domain, which may serve as a "functional unit" for retention of β4GT in the *trans*-Golgi. Also note that the $NH_2$-terminal 13 amino acids, which have been proposed to override the trans-Golgi retention signal and thereby direct some of the long mammalian β4GT protein isoform to the cell surface (40), are absent in both chicken β4GT homologues. Although it had been reported that the cytoplasmic domain of the human sequence lacks the Ser residue at amino acid 11 (8), when the human cDNA was resequenced, we found that the trinucleotide encoding this residue was present, as did Watzele and Berger (43).

In the context of the biological significance of this 13-amino acid extension, a comparison of the cytoplasmic domains of the mammalian and the two chicken β4GT proteins is instructive. Since β4GT has been reported on the cell surface of a variety of chicken cells and tissues (see Ref. 42 and references therein), one would anticipate that a functional domain responsible for the redirection of a protein from the Golgi to the cell surface would be conserved between mammals and chickens.

From an inspection of the amino acid sequences of the respective cytoplasmic domains, two features stand out. First, the 13-amino acid extension characteristic of the mammalian long β4GT protein isoform is absent in both chicken β4GT homologues. Second, in place of this 13-amino acid extension, a tetra- or pentapeptide is present, which with the exception of the initiating Met, does not have any sequence in common with the mammalian $NH_2$-terminal extension. The lack of conservation in the amino acid sequence of the cytoplasmic domains between the two chicken and the mammalian β4-galactosyltransferases needs to be taken into account when considering the functional role for the 13-amino acid extension that distinguishes the long β4GT protein isoform in mammals.

*In Contrast to the Murine β4GT Gene, Transcription of the CKβ4GT-I Gene Takes Place at a Single Start Site*—Based on a detailed promoter analysis of the murine β4GT gene, we have provided a biological and functional rationale for the unusual structure of the 5′-end of this glycosyltransferase gene (Fig. 8). Specifically, we have proposed a model of transcriptional and translational 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 (10, 11). The essential feature of our model is that mammals have evolved a two-step mechanism to generate the elevated levels of β4GT enzymatic activity, in the lactating mammary gland, that are required for lactose biosynthesis. In step one, there is an upregulation of the steady state levels of β4GT mRNA, due to increased transcription from the 3.9-kb start site. In step two, the 3.9-kb β4GT transcript is translated more efficiently, relative to its housekeeping counterpart, due to deletion of most (~200 nt) of the long GC-rich 5′-untranslated sequence characteristic of the 4.1-kb mRNA.

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 β4GT for lactose biosynthesis (11). As

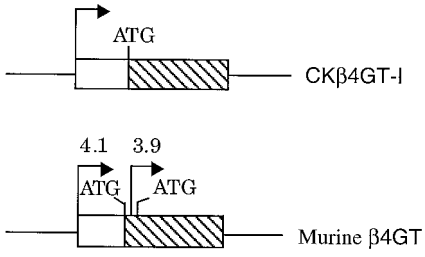


FIG. 8. **Comparison of the 5′-end of the chicken and murine β4GT genes.** The *open* and *hatched boxes* represent the 5′-untranslated and coding regions, respectively, of exon 1. The *horizontal arrows* denote the transcriptional start site relative to the position of the initiating ATG. Note that in contrast to the mammalian β4GT gene, which has two transcriptional start sites that are positioned either upstream (4.1) or between (3.9) the first two in-frame ATGs, the CKβ4GT-I gene has a single transcriptional and translational start site. This observation supports the concept that the 3.9-kb transcriptional start site, which is used preferentially in the lactating mammary gland, along with its accompanying tissue-restricted regulatory elements, have evolved in mammals to accommodate the recruited role of β4GT for lactose biosynthesis (11).

pointed out in the Introduction, a prediction of this model is that, because β4GT in nonmammalian vertebrates functions exclusively in a housekeeping (glycan biosynthesis) role, the gene will exhibit only one (or one set of) clustered transcriptional start site(s), characteristic of many housekeeping genes. The structures of the respective 5′-end of the CKβ4GT-I and the murine β4GT gene are summarized in Fig. 8. Note that in contrast to the mammalian β4GT gene, the CKβ4GT-I gene has a single transcriptional and translational start site; consequently, this prediction of our model has been substantiated. In our view, these data support the concept that these additional features of the mammalian β4GT gene were introduced into the ancestral vertebrate CKβ4GT-I gene lineage, during the evolution of mammals, as a direct consequence of the recruitment of this galactosyltransferase for the mammary gland-specific biosynthesis of lactose.

*Acknowledgments*—We thank Drs. Neng-Wen Lo, Jonathan Pevsner, and William Wright for critical commentary on this manuscript and for insightful discussions. We thank Pascale Schoenmakers for assistance with the expression of the cDNA clones.

REFERENCES

1. Beyer, T. A., and Hill, R. L. (1982) *The Glycoconjugates* (Horowitz, M., ed) pp. 25–45, Academic Press, Inc., New York
2. Brodbeck, U., Denton, W. L., Tanahashi, N., and Ebner, K. E. (1967) *J. Biol. Chem.* **242,** 1391–1397

3. Brew, K., Vanaman, T. C., and Hill, R. L. (1968) *Proc. Natl. Acad. Sci.* **59,** 491–497
4. Hill, R. L., Brew, K., Vanaman, T. C., Trayer, I. P., and Mattock, P. (1968) *Brookhaven Symp. Biol.* **21,** 139–154
5. Powell, J. T., and Brew, K. (1974) *Biochem. J.* **142,** 203–209
6. Shaper, N. L., Hollis, G. F., Douglas, J. G., Kirsch, I. R., and Shaper, J. H. (1988) *J. Biol. Chem.* **263,** 10420–10428
7. Russo, R. N., Shaper, N. L., and Shaper, J. H. (1990) *J. Biol. Chem.* **265,** 3324–3331
8. Masri, K. A., Appert, H. E., and Fukuda, M. N. (1988) *Biochem. Biophys. Res. Commun.* **157,** 657–663
9. Mengle-Gaw, L., McCoy-Haman, M. F., and Tiemeier, D. C. (1991) *Biochem. Biophys. Res. Commun.* **176,** 1269–1276
10. Harduin-Lepers, A., Shaper, N. L., and Shaper, J. H. (1993) *J. Biol. Chem.* **268,** 14348–14359
11. Rajput, B., Shaper, N. L., and Shaper, J. H. (1996) *J. Biol. Chem.* **271,** 5131–5142
12. Shaper, N. L., Shaper, J. H., Peyser, M., and Kozak, C. A. (1990) *Cytogenet. Cell Genet.* **54,** 172–174
13. Shaper, N. L., Shaper, J. H., Bertness, V., Chang, H., Kirsch, I. R., and Hollis, G. F. (1986) *Somatic Cell Mol. Genet.* **12,** 633–636
14. Shaper, J. H., Meurer, J. A., Joziasse, D. H., Chou, T.-D. D., Schnaar, R. L., and Shaper, N. L. (1995) *Glycoconjugate J.* **12,** 477–478
15. Akiyama, Y., and Kato, S. (1974) *Biken J.* **17,** 105–116
16. Langlois, J., Lapis, K., Ishizaki, R., Beard, J. W., and Bolognesi, D. P. (1974) *Cancer Res.* **34,** 1457–1464
17. Shaper, N. L., Shaper, J. H., Meuth, J. L., Fox, J. L., Chang, H., Kirsch, I. R., and Hollis, G. F. (1986) *Proc. Natl. Acad. Sci. U. S. A.* **83,** 1573–1577
18. Tessier, D. C., Thomas, D. Y., Khouri, H. E., Laliberté, F., and Vernet, T. (1991) *Gene* (*Amst.*) **98,** 177–183
19. Joziasse, D. H., Shaper, N. L., Salyer, L. S., Van den Eijnden, D. H., van der Spoel, A. C., and Shaper, J. H. (1990) *Eur. J. Biochem.* **191,** 75–83
20. Joziasse, D. H., Shaper, N. L., Kim, D., Van den Eijnden, D. H., and Shaper, J. H. (1992) *J. Biol. Chem.* **267,** 5534–5541
21. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979) *Biochemistry* **18,** 5294–5299
22. Schuler, G. D., *et al.* (1996) *Science* **274,** 540–546
23. Smith, E. J., Cheng, H. H., and Vallejo, R. L. (1996) *Poultry Sci.* **75,** 642–647
24. Kozak, M. (1984) *Nucleic Acids Res.* **12,** 857–872
25. Yadav, S., and Brew, K. (1991) *J. Biol. Chem.* **266,** 698–703
26. Yadav, S., and Brew, K. (1990) *J. Biol. Chem.* **265,** 14163–14169
27. Russo, R. N. (1990) *Two Forms of β1,4-Galactosyltransferase*, Ph.D. thesis, Johns Hopkins University
28. Ulrich, J. T., Schenck, J. R., Rittenhouse, H. G., Shaper, N. L., and Shaper, J. H. (1986) *J. Biol. Chem.* **261,** 7975–7981
29. Aoki, D., Appert, H. E., Johnson, D., Wong, S. S., and Fukuda, M. N. (1990) *EMBO J.* **9,** 3171–3178
30. Hollis, G. F., Douglas, J. G., Shaper, N. L., Shaper, J. H., Stafford-Hollis, J. M., Evans, R. J., and Kirsch, I. R. (1989) *Biochem. Biophys. Res. Commun.* **162,** 1069–1075
31. Klausner, R. D., and Rouault, T. A. (1993) *Mol. Biol. Cell* **4,** 1–5
32. Saitoh, Y., Ogawa, A., Hori, T., Kunita, R., Mizuno, S. (1993) *Chromosome Res.* **1,** 239–251
33. Nanda, I., Tanaka, T., and Schmid, M. (1996) *Gene* (*Amst.*) **170,** 159–164
34. Asano, M., Furukawa, F., Kido, M., Matsumoto, S., Umesaki, Y., Kochibe, N., and Iwakura, Y. (1997) *EMBO J.* **16,** 1850–1857
35. Lu, Q., Hasty, P., and Shur, B. D. (1997) *Dev. Biol.* **181,** 257–267
36. Colley, K. J. (1997) *Glycobiology* **7,** 1–13
37. Machamer, C. E. (1993) *Curr. Opin. Cell Biol.* **5,** 606–612
38. Shaper, J. H., and Shaper, N. L. (1992) *Curr. Opin. Struct. Biol.* **2,** 701–709
39. Russo, R. N., Shaper, N. L., Taatjes, D. J., and Shaper, J. H. (1992) *J. Biol. Chem.* **267,** 9241–9247
40. Lopez, L. C. Youakim, A., Evans, S. C., and Shur, B. D. (1991) *J. Biol. Chem.* **266,** 15984–15991
41. Appeddu, P. A., and Shur, B. D. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91,** 2095–2099
42. Hathaway, H. J., and Shur, B. D. (1992) *J. Cell Biol.* **117,** 369–382
43. Watzele, G., and Berger, E. G. (1990) *Nucleic Acids Res.* **18,** 7174