Immunogenetics (1999) 49:420–428

ORIGINAL PAPER

**Moon G. Kim · Chuan Chen · Myung Soo Lyu
Eun-Gyung Cho · Dongeun Park · Christine Kozak
Ronald H. Schwartz**

Cloning and chromosomal mapping of a gene isolated from thymic stromal cells encoding a new mouse type II membrane serine protease, epithin, containing four LDL receptor modules and two CUB domains

Received: 6 May 1998 / Revised: 1 September 1998

Abstract We cloned and sequenced a mouse gene encoding a new type of membrane bound serine protease (epithin) containing a multidomain structure. The initial cDNA clone was found previously in a polymerase chain reaction (PCR)-based subtractive library generated from fetal thymic stromal cells, and the message was shown to be highly expressed in a thymic epithelial nurse cell line. A clone isolated from a severe combined immunodeficiency (SCID) thymus library and extended to its full length at the 5' end with the RACE technique contains an open reading frame of 902 amino acids. Based on the sequence of this clone, the predicted protein structure is a type II membrane protein with a C-terminal serine protease domain linked to the membrane by four low density lipoprotein receptor modules and two CUB domains. High message expression by northern blotting was detected in intestine, kidney, lung, SCID, and Rag-$2^{-/-}$ thymus, and 2-deoxyguanosine-treated fetal thymic rudiment, but not in skeletal muscle, liver, heart, testis, and brain. Sorted MHC class II^+ and II^- fetal thymic stromal cells were positive for expression by reverse transcriptase-PCR, whereas $CD45^+$ thymocytes were not. The gene was found in chicken and multiple mammalian species under low stringency Southern hybridization conditions. Under high stringency conditions, only a single gene per haploid genome was identified in the mouse. This gene, *Prss14* (*protease*, *serine*, *14*), was mapped to mouse chromosome 9 and is closely linked to the *Fli1* (Friend leukemia integration 1) gene.

Key words Thymic stroma · Serine protease · CUB domain · LDLR domain · Type II membrane protein

M.G. Kim · C. Chen · R.H. Schwartz (✉)
Laboratory of Cellular and Molecular Immunology, National Institute of Allergy and Infectious Diseases,
National Institutes of Health, Bethesda MD 20892-0420 USA,
E-mail: rschwartz@atlas.niaid.nih.gov,
Tel.: +1-301-496-1257, Fax: +1-301-496-0877

M.S. Lyu · C. Kozak
Laboratory of Molecular Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda MD 20892-0420, USA

E.-G. Cho · D. Park
Department of Molecular Biology, Seoul National University, Seoul, Korea

Introduction

Thymic stromal cells, especially thymic epithelial cells, play important roles in thymopoiesis, and they produce signals for thymocyte survival, migration, selection, and death (Boyd and Hugo 1991; Boyd et al. 1993; van Ewijk 1991). Our knowledge of the expressed genes associated with thymic stromal function is rather limited. One of the few known useful examples is the *nude* mutation (Nehls et al. 1994, 1996). Mice with this mutation lack a thymus because of a defect in epithelial cell function at an early stage of development (fetal day 11). In order to generate more molecular reagents to study mouse thymic stromal cells, we previously generated a polymerase chain reaction (PCR)-based subtractive cDNA library from isolated fetal thymic stromal cells (Kim et al. 1998). Among the newly identified genes, *TSO-3.39* (TSO for thymic stromal origin) was expressed in the severe combined immunodeficiency (SCID) thymus and a thymic epithelial cell line corresponding to a nurse cell, but not in several other types of thymic epithelial or fibroblast cell lines. In this paper we present the sequence of the full-length cDNA, a more extensive tissue expression pattern, the predicted features of the protein, and the chromosomal mapping.

Materials and methods

Cells, tissues, and mice

SV40-transformed thymic epithelial cell lines (Faas et al. 1993) from transgenic animals expressing the SV40 T antigen were the kind gift of B. Knowles, Jackson Laboratory, Bar Harbor, Me. The Tst1 thymic epithelial cell line derived from a P53 knock-out mouse was the generous gift of M. Sitkovsky, LI, NIAID, NIH, Bethesda, Md. The NIH3T3 and AKR6.1 cell lines were obtained from the ATCC (Rockville, Md.). The 2B4 T-cell hybridoma was developed in our laboratory by L. Samelson (Samelson and Schwartz 1983).

Fetal thymii of 14.5 days gestation were obtained from timed matings of C57BL/6 (B6) mice (NCI, Frederick, Md.). The day of appearance of the vaginal plug was counted as day 0. C.B-17 mice bearing the SCID mutation were bred in our animal facility (NIAID, Frederick, Md.). Mice bearing the RAG-$2^{-/-}$ targeted mutation were obtained from the laboratory of F. Alt (Shinkai et al. 1992) and derived into our breeding colony at Taconic Farms, Inc. by embryo transfer. They were backcrossed onto the C57BL/10 background and used at generation N11. Crude thymic stroma from adult thymuses of normal or SCID animals were prepared by washing the thymocytes away from the fibrous stromal tissue after dispersing and filtering the organ through a nylon membrane. Purified thymic stromal cells and $CD45^+$ thymocytes were sorted in the NIAID flow cytometry unit from fetal thymic organ cultures (FTOC), treated or not with 2-deoxyguanosine, using anti-I-A^b PE (AF6-120.1, Pharmingen, San Diego, CA) and anti-CD45 FITC (30F11.1, Pharmingen) (Kim et al. 1998). Purity of the sorted cell populations was 99.4% for $CD45^+$ cells and greater than 94% for the MHC II^+ stromal cells.

Cloning of the full length cDNA

A 0.4 kilobase (kb) cDNA clone (TSO3.39) from a PCR-based subtractive thymic stromal cell cDNA library was used to screen another cDNA library prepared from thymuses of mice with the severe combined immunodeficiency (SCID) mutation (protein kinase, DNA activated, catalytic polypeptide mutation, *Prkdc*^{scid}). Among 15 isolated clones, the longest cDNA from the SCID thymus library was 1.8 kb (pBK3.39.11b, inserted at the *Eco* RI and *Xho* I sites of the vector), but it was still missing the 5' end of the cDNA. The 5' end (1.5 kb) was generated by the rapid amplification of cDNA end (RACE) method from an adaptor-ligated SCID thymus cDNA pool generated by using a Marathon cDNA Amplification Kit (Clontech, Palo Alto, CA). RACE76, one of four independent clones, was then cut and pasted at a unique *Bcl* I site which generates a 75 base pair (bp) overlap between the two clones. Oligonucleotide sequences used for the RACE primers were TGGCATTGCATCGGCAGTAAC (339 Race6) and CACGGGTCGTTGGAGTCGTAG (339 Race5). The amplification was carried out for 25 cycles at 94 °C for 30 s and 68 °C for 4 min using a mixture of *Taq* and *pfu* polymerases (Stratagene, La Jolla, CA).

DNA sequencing

Cycle sequencing was performed using the ABI prism dye-terminator sequencing kit according to the standard protocol with 0.4 μg of plasmid DNA prepared with a Qiagen miniprep kit (Qiagen, Valencia, CA) using T3 and T7 sequencing primers, or internal primers.

Database search, primary sequence analysis, and sequence alignment

The sequence similarity searches (BLAST) of the nonredundant data base managed by NCBI were carried out using Blast Client 1.4 (W. Gilbert of the Whitehead Institute for Biomedical Research) or using the Blast Enhanced Utility Program (BEUTY) at the BCM server (http://kiwi.imgen.bcm.tmc.edu). Multiple nucleotide sequence assembly was performed using AssemblyLIGN 1.0 (IBI). ClustalW alignments of multiple amino acid sequences were done starting with the PAM or BLOSUM programs in MacVector 6.0 (Oxford Molecular Group, Campbell, CA). Codon preference profile was checked with CodonUse 3.5.3 developed by Dr. Conrad Halling (chhall@bb1t.monsanto.com). The membrane topology was analyzed using TopPred II (Claros and von Heijne 1994) and PSORT at the Nakai server (http://psort.nibb.ac.jp).

Northern blot analysis

Total RNAs from mouse tissues were prepared using Trizol (Molecular Research Center, INC, Cincinnati, OH) according to the manufacturer's instructions. Poly A^+ mRNAs were prepared using a FastTrack 2.0 kit (Invitrogen, Carlsbad, CA). Ten micrograms of total RNA or 2 μg of polyA$^+$ mRNA was electrophoresed on a formaldehyde containing 1.0% agarose gel and blotted onto a nylon membrane, Nytran maximum strength (Schleicher and Schuell, Keene, NH). The membrane was hybridized with a [α-^{32}P] dATP-labeled probe (the 1.8 kb *Eco* RI to *Xho* I fragment of pBK3.39.11b) in QuikHyb solution (Stratagene) for 2 hours and washed twice with 2× SSC-0.1% SDS, once with 0.1× SSC-0.1% SDS, each time at 65 °C for about 20 minutes, and then analyzed with a phosphor imager (Molecular Dynamics, Sunnyvale, CA). The experiment in Fig. 1B was done with a multiple tissue Northern blot membrane prepared with polyA$^+$ mRNA from mice (Clontech).

Reverse transcriptase-polymerase chain reaction

MHC class II^+ and MHC class II^- thymic stromal cells were prepared by sorting cell suspensions made from 2-deoxyguanosine-treated fetal thymic organ cultures (FTOC) (Kim et al. 1998). $CD45^+$ thymocytes were prepared by sorting cells from parallel FTOC not treated with 2-deoxyguanosine. Total RNA was prepared from sorted cells and resuspended at a concentration equivalent to 1000 cells/μl. Twenty-five microliters of each of these RNAs or 10 μg of total RNA from either deoxyguanosine-treated thymus or SCID thymus were reverse transcribed using MMTV RT (Stratagene) in a total volume of 50 μl. One microliter of each generated cDNA was then amplified with Tag2000 (Stratagene) using epithin-specific primers (Forward 5' primer #6: nucleotides 2432–2454, AGAAGGGTGAGATCCGTGTCATC and Backward 3' primer #22: nucleotides 2902–2878, AAGTTCTCTCTC-CAACTCTTGAGGG). The sequences of primers used for the GAPDH controls are GGTGAAGGTCGGTGTGAACGGA for the 5' primer, and TGTTAGTGGGGTCTCGCTCCTG for the 3' primer.

Southern blot analysis

A Zoo-Blot (Clontech) containing *Eco* RI-digested genomic DNAs from eight eukaryotic species: human, monkey, rat, mouse, dog, cow, rabbit, and chicken was probed with the whole insert (1.8 kb *Eco* RI to *Xho* I fragment) of pBK3.39.11b. For the low stringency analysis, the filter was washed in 2× standard sodium citrate (SSC)-0.1% sodium dodecyl sulfate (SDS) at 65 °C, and for the high stringency analysis, in 0.1× SSC-0.1% SDS at 65 °C.

Chromosomal localization

TSO-3.39 was mapped by analysis of two sets of genetic crosses: (NFS/N or C58/J X *M. m. musculus*) X *M. m. musculus* (Kozak et al. 1990), and (NFS/N X *M. spretus*) X *M. spretus* or C58/J (Ad-

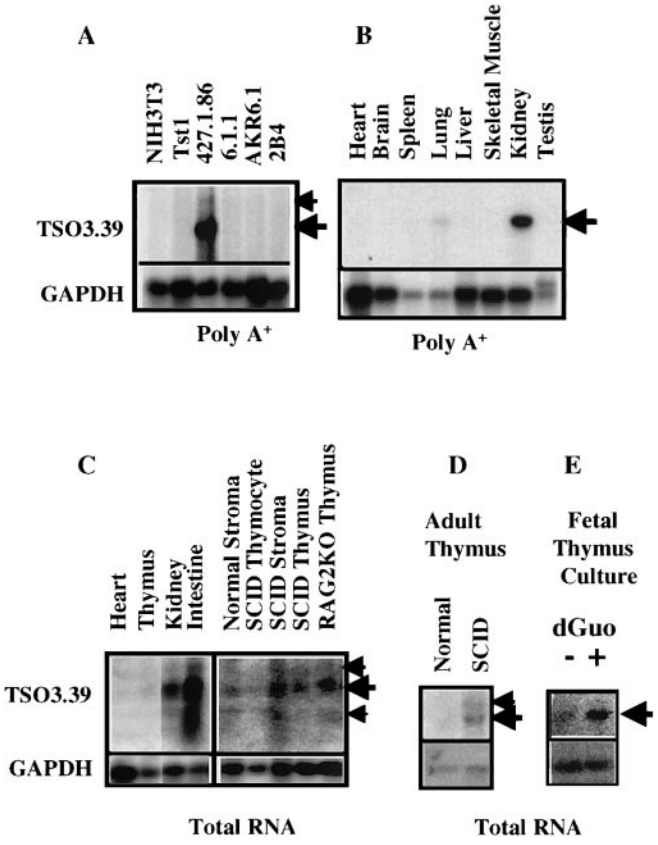


Fig. 1A–E The mRNA expression pattern in different cell lines and tissues with the TSO-3.39 probe. The northern blots were carried out with 2 μg of polyA$^+$ RNA or 10 μg of total RNA from the tissues indicated at the *top* of each gel. *Big arrows* indicate the major message, and *small arrows* indicate occasional additional bands that were observed. For all experiments, GAPDH was used to normalize for the amount of RNA loaded. All the samples except testis showed a single GAPDH message. **A** Expression is restricted to a thymic nurse cell line. PolyA$^+$ RNA from NIH3T3 (fibroblast line), Tst1 (a thymic epithelial cell line derived from a p53 knock-out mouse), 427.1.86 (thymic nurse cell line derived from an SV40 T-antigen transgenic mouse), 6.1.1 (thymic medullary epithelial line from the same transgenic mouse), AKR6.1 (a thymoma line with thymocyte characteristics), and 2B4 (a T-cell hybridoma) were used. **B** Expression in various tissues from adult mouse. A multi-tissue northern blot membrane (Clontech) is shown containing 2 μg of polyA$^+$ RNA from the tissues listed *above* the blot. **C** Expression in more tissues and various fractions of SCID, Rag-2$^{-/-}$, or normal thymus using total RNA. The crude thymic stroma were prepared by harvesting the cells remaining after removing thymocytes through a nylon mesh. **D** Comparison of the expression of RNA from thymuses of normal and SCID mice. The TSO-3.39 message can also be detected in the RNA from normal thymus, but only after a longer exposure. This is not shown because it interferes with the identification of the two bands in the SCID thymus. **E** RNA from fetal thymus organ cultures with or without 2-deoxyguanosine (dGuo)

amson et al. 1991). DNAs from the progeny of these crosses were typed for over 1000 markers which map to all 19 autosomes and the X chromosome.

Data were stored and analyzed using the program LOCUS. Percent recombination and standard errors between specific loci were calculated from the number of recombinants according to Green (1981). Loci were ordered by minimizing the number of recombinants.

Results and discussion

Among the genes from a PCR-based subtractive thymic stromal cell cDNA library generated from the cells of deoxyguanosine-treated fetal thymic organ cultures, a 0.4 kb clone, TSO-3.39, showed an interesting expression pattern when tested by northern blot analysis (Kim et al. 1998). In the initial screening, we used five transformed thymic stromal cell lines (Faas et al. 1993), as well as two fibroblast cell lines, in order to identify genes that potentially were specifically expressed in subtypes of thymic stromal cells. TSO-3.39 expression was detected only in the nurse cell type thymic epithelial cell line (427.1.86 cell) on a northern blot using total RNA. Figure 1 A shows a similar analysis done on a northern blot with 10-fold-enriched poly A$^+$ RNA. Again mRNA expression was restricted to the thymic nurse cell line. No message was detected in a fibroblast line (NIH3T3), other types of thymic epithelial cell lines (Tst1 and 6.1.1), a T-cell thymoma (AKR6.1) or a T-cell hybridoma (2B4). Overall, the limited expression pattern suggested that the gene might play a unique role in this specialized epithelial compartment in the thymus. Nurse cells are known to be involved in thymic selection (Lahoud et al. 1993). In particular, the 427.1.86 cell line has been shown to function in the positive and negative selection of T cells when injected intrathymically (Vukmanovic et al. 1992; Vukmanovic et al. 1994a, 1994b).

A more extensive analysis of the steady state level of TSO-3.39 message expression in different tissues was performed by northern blotting using either poly A$^+$ RNA or total RNA (Fig. 1B-E). TSO-3.39 expression was most prominent in the intestine and the kidney, and a weak band was observed with RNA from the lung (Fig. 1B, C). No bands were detected with RNA from the brain, heart, liver, testis, or skeletal muscle. Thymus and spleen gave very weak signals, which were clearly visible on longer exposure (data not shown). By contrast, RNA from SCID and Rag-2$^{-/-}$ thymuses showed strong expression of TSO-3.39 (Fig. 1C, D). These mice lack fully developed T and B cells (Bosma and Carroll 1991; Bosma et al. 1983, 1988), because of a defect in the joining of their antigen-specific receptor genes (Kim et al. 1988; Kirchgessner et al. 1995; Lieber et al. 1988). Thus, their thymuses would be enriched in stromal cell content compared to a normal thymus.

To more directly assess which cells in the thymus are expressing TSO-3.39, we fractionated SCID thymus into a thymocyte cell suspension and stromal cells. RNA from the stromal cells showed a strong band on northern blotting, while RNA from the thymocytes showed only a weak band (Fig. 1C). Stromal cells isolated from normal adult thymus also expressed TSO-

3.39. In addition, expression was about 10-fold more in the thymic rudiment generated from FTOC treated with 2-deoxyguanosine compared to organ cultures not treated with the drug (Fig. 1E). 2-deoxyguanosine depletes thymocytes, leaving only thymic stromal cells. It was the source of our original subtractive cDNA library**.** These data support the idea that TSO-3.39 expression is limited to the thymic stromal compartment.

The size of the major message on the northern blots is approximately 3 kb. It should be noted, however, that there are extra bands in some tissues with different levels of expression of each form. In the intestine, SCID thymic stroma, and Rag-$2^{-/-}$ thymus, there are additional smaller and larger bands, while in the thymic nurse cell line 427.1.86 and in the SCID thymus there is only an additional larger band. These different forms of the mRNA could result from alternative splicing or be due to alternative polyadenylation sites. These possibilities remain to be investigated.

Reverse transcriptase-polymerase chain reaction detection of TSO-3.39 expression in thymus subpopulations

In order to verify that TSO-3.39 was expressed in thymic stromal cells and not in $CD45^+$ thymocytes, we isolated these cell populations by flow cytometry from FTOC, after staining with MHC-specific class II and CD45-specific antibodies (Kim et al. 1998), and tested the purified cells for expression of TSO-3.39 by reverse transcriptase (RT)-PCR. As shown in Fig. 2, message was clearly detected in the MHC class II^+ and MHC class II^- thymic stromal cells, but not in $CD45^+$ thymocytes. The amount of RNA used in this RT-PCR reaction was isolated from about 500 cells. The positive controls show the signal from 0.2 μg of total RNA prepared from deoxyguanosine-treated FTOC or SCID thymic tissue. We conclude from this experiment that thymic stromal cells of at least two different types express TSO-3.39 mRNA.

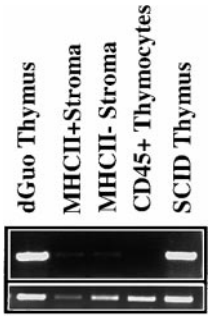


Fig. 2 TSO-3.39 expression detected by RT-PCR in fractionated thymic cells. Single-cell suspensions of MHC class II^+ and MHC class II^- stromal cells, and $CD45^+$ thymocytes were prepared by FACS sorting. The results with 0.2 μg of total RNA from dGuo-treated thymic rudiments and SCID thymus are shown as positive controls. The PCR product from GAPDH message is shown at the *bottom*

```
   1 MGSNRGRKAG  GGSQDFGAGL  KYNSRLENMN  GFEEGVEFLP  ANNAKKVEKR
     GPRRWTVLVA  VLFSFLLLSL  MAGLLVWHFH  YRNVRVQKVF  NGHLRITNEI
 101 FLDAYENSTS  TEFISLASQV  KEALKLLYNE  VPVLGPYHKK  SAVTAFSEGS
     VIAYYWSEFS  IPPHLAEEVD  RAMAVERVVT  LPPRARALKS  FVLTSVVAFP
 201 IDPRMLQRTQ  DNSCSFALHA  HGAAVTRFTT  PGFPNSPYPA  HARCQWVLRG
     DADSVLSLTF  RSFDVAPCDE  HGSDLVTVYD  SLSPMEPHAV  VRLCGTFSPS
 301 YNLTFLSSQN  VFLVTLITNT  GRRHLGFEAT  FFQLPKMSSC  GGVLSDTQGT
     FSSPYYPGHY  PPNINCTWNI  KVPNNRNVKV  RFKLFYLVDP  NVPVGSCTKD
 401 YVEINGEKGS  GERSQFVVSS  NSSKITVHFH  SDHSYTDTGF  LAEYLSYDSN
     DPCPGMFMCK  TGRCIRKELR  CDGWADCPDY  SDERYCRCNA  THQFTCKNQF
 501 CKPLFWVCDS  VNDCGDGSDE  EGCSCPAGSF  KCSNGKCLPQ  SQKCNGKDNC
     GDGSDEASCD  SVNVVSCTKY  TYRCQNGLCL  SKGNPECDGK  TDCSDGSDEK
 601 NCDCGLRSFT  KQARVVGGTN  ADEGEWPWQV  SLHALGQGHL  CGASLISPDW
     LVSAAHCFQD  DKNFKYSDYT  MWTAFLGLLD  QSKRSASGVQ  ELKLKRIITH
 701 PSFNDFTFDY  DIALLELEKS  VEYSTVVRPI  CLPDATHVFP  AGKAIWVTGW
     GHTKEGGTGA  LILQKGEIRV  INQTTCEDLM  PQQITPRMMC  VGFLSGGVDS
 801 CQGDSGGPLS  SAEKDGRMFQ  AGVVSWGEGC  AQRNKPGVYT  RLPCSSGLDQ
     RAHWGIAAWT  DSRPQTPTGM  PDMHTWIQER  NTDDIYAVAS  PPQHNPDCEL
 901 HP
```

Fig. 3 Primary amino acid sequence of the open reading frame. The predicted transmembrane region (amino acids 56–76) is *boxed* with *double lines*. The area of the LDLR domains is *boxed* with a *dotted line*. The area of the serine protease domain is *boxed* with a *dark line*. The active site motifs in the serine protease domain are *underlined* with *dotted lines*. Potential sites for N-linked sugars are shown as an *underlined* N (N107, N302, N365, N421, N489, and N772)

The full-length cDNA sequence

The completed full-length clone, pBK3.39full$^+$, contained a 3172 bp insert with an open reading frame of 902 amino acids (Fig. 3). The putative initiating methionine codon starts at nucleotide 127 and the termination codon ends at 2833 (Genebank accession number AF042822). The Kozak consensus sequence for eukaryotic translation initiation ACCATGG is present at the predicted initiation codon. The open reading frame was further verified by scanning the whole nucleotide sequence for codon preferences using CodonUse 3.5.3 (personal communication, chhall@bb1t.monsanto.com). We will refer to the protein encoded by this cDNA as epithin.

Protein primary amino acid sequence

The longest open reading frame deduced from the nucleotide sequence encodes a glycoprotein of 902 amino acids with six predicted N-linked glycosylation sites (N107, N302, N365, N421, N489, N772) (Fig. 3). The calculated relative molecular mass of the predicted protein (epithin) is about 100 000 M_r, and the estimated pI of the core protein is 6.18.

To determine whether the gene encoded a membrane protein and, if so, what was its topological nature, the complete open reading frame was analyzed at the PSORT server (see Materials and methods). Epithin was predicted to be a type II plasma membrane protein. Another program used was TopPred II (Claros

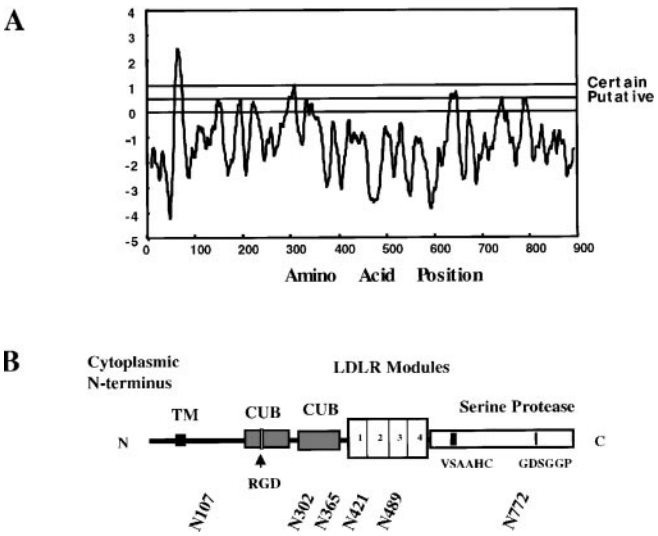


Fig. 4A,B Structural analysis of the protein Epithin. **A** A diagram of the membrane topology analysis using TopPred II. **B** Predicted domain structure of epithin. N-linked glycosylation sites are indicated at the *bottom*. *TM*, transmembrane domain; *CUB*, CUB domains; *RGD*, a single arg, gly, asp sequence; *LDLR Modules*, low density lipoprotein receptor ligand binding repeats (1–4); *Serine protease*, serine protease domain with the two conserved catalytic sites shown as *black boxes*

and von Heijne 1994); it gave a similar result. The accuracy of TopPred II is 95% for prokaryotic proteins and 83% for eukaryotic proteins. As shown in Fig. 4A, the amino acid sequence contained a single definite ("certain") transmembrane region spanning residues 56–76. This program also predicts that the 55 amino acid N terminus of the protein is inside the cell membrane, i.e., epithin is a type II membrane protein.

The mechanism of insertion into the membrane for these proteins is not well characterized. A large proportion of leukocyte cell surface proteins with enzymatic activity in their extracellular domains are type II membrane proteins (Barclay et al. 1997). Among them, CD10, CD13, CD26, and BP-1/6C3 are known to be proteases. All of these known proteases are also expressed in epithelial cells of the kidney and gut in addition to thymic stromal cells. Several of these proteins were shown to be expressed on thymic stromal cells that interact with double positive thymocytes (Small et al. 1996).

Multidomain structure of epithin

A database search for sequence similarities using predicted amino acid sequences revealed significant matches with multiple different proteins. Depending on the parts of the protein sequence put into the search, the closest matches were quite different. At first glance, they seemed to be unrelated. The C-terminal region matched with known serine proteases such as enterokinase (the smallest sum probability of the BLAST search was $9.9e^{-59}$) and mast cell protease 6 ($1.3e^{-58}$). The membrane proximal extracellular domain matched with metalloprotease family members such as the bone morphogenic protein 1 ($2.7e^{-14}$) and the Drosophila tolloid homologue ($1.3e^{-11}$). The middle portion of the extracellular region matched with LDL receptor family members including megalin ($1.4e^{-43}$). Therefore, we divided epithin into five separate domains: the cytoplasmic, N-terminal domain (amino acids 1–55), the transmembrane domain (amino acids 56–76), the membrane proximal extracellular region containing 2 CUB domains (amino acids 77–450), the region containing the LDL receptor ligand binding modules (amino acids 451–600), and the C-terminal serine protease domain (amino acids 601–902) as shown in Figs. 3, 4B.

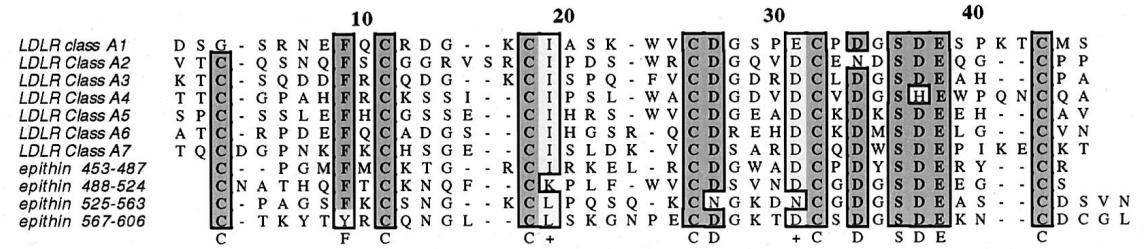
In the putative protease domain there are two conserved amino acid sequences (VSAAHC and GDSGGP shown in Figs. 3, 4B). All members of the trypsin family of serine proteases contain these sequences, except for a few cases in which the first glycine is missing (Prosite, PDOC00124, Rawlings and Barrett 1994). If a protein contains both the serine (GDSGGP) and the histidine (VSAAHC) active site motifs in the catalytic domain, the probability of it being a trypsin family serine protease is 100% (Prosite, PDOC00124).

The putative LDL receptor modules found in epithin are compared in Fig. 5A with the seven known LDL receptor ligand binding modules from the LDL receptor. The six conserved cysteines (positions C3, C11, C18, C26, C32, and C43 in Fig. 5A) are all present in the epithin repeats. The amino acids that are known to be involved in the coordination of calcium (D27, D31, D37, and E38 in Fig. 5A) are also present in their conserved positons. A recent crystallographic study of the LDL receptor has suggested the importance of these conserved cysteines and acidic residues for proper folding and lipoprotein binding (Fass et al. 1997).

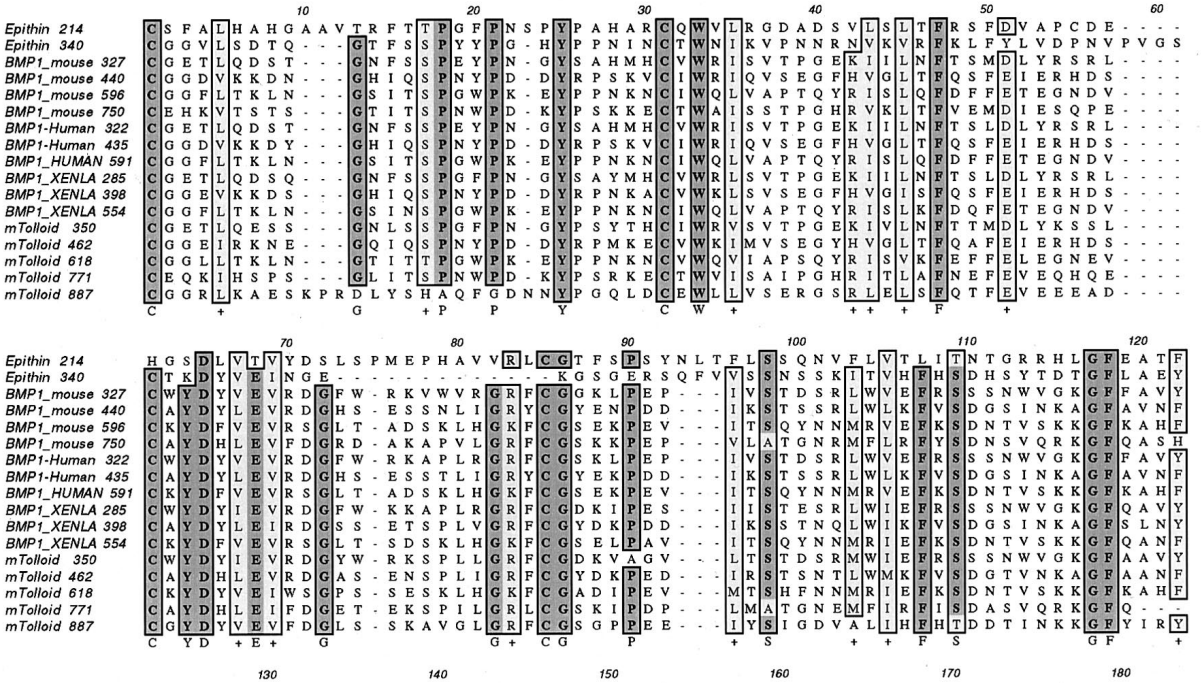
The membrane proximal extracellular region contains two conserved amino acid blocks called CUB domains (Prosite, PDOC00908) which we identified in a

Fig. 5 A ClustalW formatted amino acid alignments of the LDLR ligand binding modules in the LDL receptor and epithin starting with the BLOSUM program. **B** and **C** ClustalW formatted amino acid alignments of the CUB domains. **B** Comparison of the epithin domains with the bone morphogenic protein 1 family and **C** comparison with seven proteases containing CUB domains. For **B** the sources of the protein sequences are: bone morphogenic protein 1 (BMP1) from mouse, human, and xenopus, as well as the mammalian tolloid homologue from mouse. For **C** the sequences are from: enterokinase from mouse, Ra-reactive factor serine protease (CRAR) from human and mouse, complement proteins C1R and C1S from human, the Ca^{++}-dependent serine protease (CASP) from hamster, which appears to be a homologue of C1S, and blastula protease 10 from sea urchin (BP-10). The starting number of the amino acid in the putative domain is shown *next to* the name of the protein. *Dark shading* depicts positions where the amino acid is >85% identical. *Light shading* depicts conservation of chemically similar amino acids

A


10
20
30
40
LDLR class A1
LDLR Class A2
LDLR Class A3
LDLR Class A4
LDLR Class A5
LDLR Class A6
LDLR Class A7
epithin 453-487
epithin 488-524
epithin 525-563
epithin 567-606

B


Epithin 214
Epithin 340
BMP1_mouse 327
BMP1_mouse 440
BMP1_mouse 596
BMP1_mouse 750
BMP1-Human 322
BMP1-Human 435
BMP1_HUMAN 591
BMP1_XENLA 285
BMP1_XENLA 398
BMP1_XENLA 554
mTolloid 350
mTolloid 462
mTolloid 618
mTolloid 771
mTolloid 887

C

Epithin 214
Epithin 340
ENT Kinase Mouse 569
CRAR Human 11
CRAR Mouse 16
C1R Human 13
C1S Human 3
CASP Hamster 9
CRAR Human 185
CRAR Mouse 190
C1R Human 193
C1S Human 175
CASP Hamster 181
BP-10 Sea Urchin 339
BP-10 Sea Urchin 484

Blocks data base search (Blksort server at http://www.blocks.fhcrc.org/blocks). The CUB domain was previously defined as an extracellular domain of approximately 110 residues that is found in functionally diverse, mostly developmentally regulated proteins and that contains two pairs of conserved cysteines (Bork 1991; Bork and Beckmann 1993). The name CUB is an acronym formed from the first three identified proteins of this family: the complement components C1R/C1S, an embryonic sea urchin protein, Uegf, and bone morphogenic protein 1, BMP-1. Because our initial blast search had identified high similarity for this region with bone morphogenic protein 1, the two CUB domain sequences (214–331, and 340–444) in epithin were first compared with the CUB domains in the proteins of the BMP-1 family by a ClustalW analysis. CUB domains aligned with the first cysteine as 1, showed 20 conserved (i.e., ≥85% of the residues were the same amino acid as in one of the epithin domains) and 13 chemically conserved residues (Fig. 5B). The epithin CUB domains appear to have diverged somewhat in the C-terminal half of the domain. In particular, one of the two cysteines forming the second half of the CUB domain is either located in a different place or missing. The same is true of other residues such as the conserved Y64 in the CxYD sequence motif.

CUB domains are also found in a variety of other proteins. Figure 5C shows a comparison of the epithin domains with the CUB domains found in seven known proteases. In this case 18 of the residues are highly conserved and 12 are chemically conserved. Thirteen of the conserved residues are found in both the BMP-1 and serine protease families. Unique to the BMP-1 family are the two motifs, CxW(33) and EV/IxxG(72), while highly represented in the protease family are a P32 following the conserved PxxPx(x)xY motif and a D121 at the end of the FxS/TD motif. Epithin, of course, has at least one copy of each of these motifs. The only conserved residues that it lacks are G46 in the protease family and G82 in the BMP-1 family.

The conserved PxxPx(x)xY motif is also found in other CUB domains from proteins without known protease functions such as ebnerine (A57190), CRP-ductin (U37438), and neuropilin (AF016297). The function of these conserved residues, as well as the CUB domain itself, is still unknown. Finally, the first CUB domain in epithin (214) contains one RGD sequence (residues 42–44 in Fig. 5C), a motif originally identified as a cell attachment site for extracellular matrix proteins via membrane molecules such as integrins. (Ruoslahti 1996). The RGD sequence is not present in any of the other CUB domains we analyzed and its functional significance remains to be tested.

The multidomain structure of epithin indicates that the gene may have evolved by exon shuffling, like many proteases with multiple domains: for example, enterokinase and bone morphogenic protein 1 (Bork and Beckmann 1993; Rawlings and Barrett 1994). The combination of seemingly unrelated domains suggests that this protein might have several biological or biochemical functions in epithelial cells. The high similarity between epithin and the enterokinase in both the CUB domains and the protease domains suggests that these two domains might be functionally coupled.

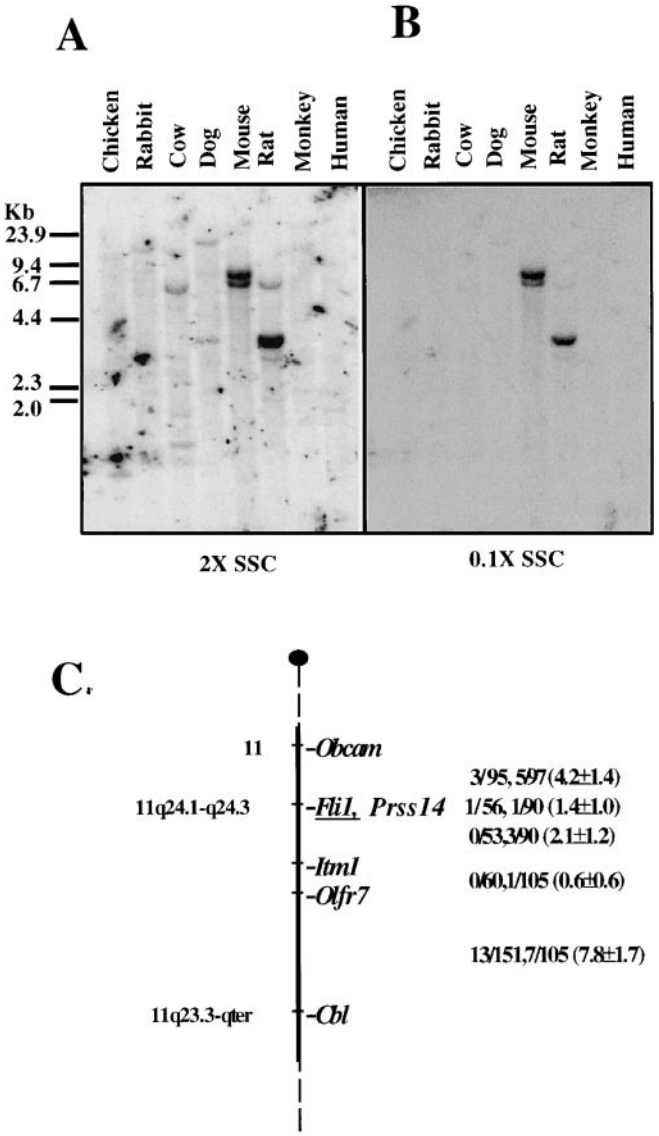


Fig. 6A–C Analysis and mapping of the gene**.** Genomic Southern blot using the 0.4 kb TSO-3.39 probe with *Eco* RI-digested DNA from different animal species (**A** and **B**). The species are indicated at the *top*. In **A** the blot was washed in 2× SSC at 65 °C (low stringency). In **B** the blot was washed in 0.1× SSC at 65 °C (high stringency). **C** Location of the *Prss14* gene on mouse Chromosome 9. Recombination fractions are given on the *right* for each locus pair, with the first fraction representing results from the *M. m. musculus* crosses and the second from the *M. spretus* crosses. Percent recombination and standard errors are given in *parentheses* and were determined according to Green (1981). Chromosome 9 marker loci were typed as previously described (Hong et al. 1996; Sullivan et al. 1996). Both of the recombinants noted between *Fli1* and *Prss14* are double recombinants around *Prss14* and, therefore, the relative positions of the two genes could not be determined. Human map locations for homologues of the *underlined* genes are given on the *left* of the map

Southern blot analysis

An *Eco* RI digest of genomic DNAs from various organisms showed that homologs of the mouse gene encoding epithin (*Prss14*, for protease, serine family member 14) are present in other species (Fig. 6). Southern blot analysis with a low stringency wash revealed multiple hybridized bands with DNA from all the organisms (human, monkey, rat, mouse, dog, cow, rabbit, and chicken) (Fig. 6 A). In mouse and rat, the two bands observed are very strong. These results indicate the existence of family members related to the *Prss14* gene in birds as well as in other mammals. In contrast, only two bands remained hybridized in mouse and rat at high stringency wash conditions (Fig. 6B), indicating that the *Prss14* gene nucleotide sequence is highly conserved between these two rodents. Digestion of DNA with other enzymes (*Pvu* II, *Sca* I, *Taq* 1, *Hpa* I, *Bgl* II, *Bam* HI, *Xba* I, *Sst* I) and hybridization with a shorter probe revealed only one band under high stringency conditions (data not shown), demonstrating that a single gene per haploid genome is present in the mouse.

Genetic mapping

The 0.4 kb probe (TSO-3.39) identified *Pvu* II fragments of 6.4 kb in NFS/N and C58/J and 3.9 kb in *M. m. musculus*, as well as *Bgl* II fragments of 5.5 kb in NFS/N and 4.6 in *M. spretus*. Some of the *M. m. musculus* cross progeny DNAs were also typed using *Sst* I which identified fragments of 7.8 and 5.2 in NFS/N and 7.1 and 5.2 in *M. m. musculus*. Inheritances of the variant fragments were scored in the progeny of both sets of crosses, and linkage was detected to markers on Chr 9 (Fig. 6C). Closest linkage was found with the marker Friend leukemia integration 1 (*Fli1*) which has been placed at 16 cM from the centromere on the mouse chromosome 9 composite map (Imai 1997) The human *Fli1* homologue, *FLI1*, maps to 11q24.1-q24.3 suggesting a location for the human *PRSS14* gene.

In conclusion, we identified a new gene expressed in epithelial cells of the thymus and several other tissues which appears to encode a new type II membrane serine protease (epithin) linked to the membrane by CUB and LDLR domains. We suspect it will target either an extracellular matrix protein or another membrane bound protein on the same or neighboring cells. In the latter case, epithin cleavage products could activate target cells, facilitating their differentiation, migration, or function. Thus, in order to understand the precise immunological role played by this protease in the thymus, future studies will be directed toward identifying its specific substrate(s).

Acknowledgments We are grateful to Dr. Luciano D'Adamio, Dr. Lynda Chiodetti, and Dr. Luca Pellegrini for reading this manuscript and making very useful suggestions for its improvement. We also thank Dr. Charles Chu for introducing us to the codon usage program.

References

Adamson MC, Silver J, Kozak CA (1991) The mouse homolog of the Gibbon ape leukemia virus receptor: genetic mapping and a possible receptor function in rodents. Virology 183:778–781

Barclay AN, Brown MH, Law SKA, Tomlinson MG, van der Merwe PA (1997) The leukocyte antigen. Academic Press, San Diego

Bork P (1991) Complement components C1r/C1s, bone morphogenic protein 1 and Xenopus laevis developmentally regulated protein UVS. 2 share common repeats. FEBS Lett 282:9–12

Bork P, Beckmann G (1993) The CUB domain. A widespread module in developmentally regulated proteins. J Mol Biol 231:539–545

Bosma GC, Custer RP, Bosma MJ (1983) A severe combined immunodeficiency mutation in the mouse. Nature 301:527–530

Bosma M, Schuler W, Bosma, G (1988) The scid mouse mutant. Curr Top Microbiol Immunol 137:197–202

Bosma MJ, Carroll AM (1991) The SCID mouse mutant: definition, characterization, and potential uses. Annu Rev Immunol 9:323–350

Boyd RL, Hugo P (1991) Towards an integrated view of thymopoiesis. Immunol Today 12:71–79

Boyd RL, Tucek CL, Godfrey DI, Izon DJ, Wilson TJ, Davidson NJ, Bean AG, Ladyman HM, Ritter MA, Hugo P (1993) The thymic microenvironment. Immunol Today 14:445–459

Claros MG, von Heijne G (1994) TopPred II: an improved software for membrane protein structure predictions. Comput Appl Biosci 10:685–686

Faas SJ, Rothstein JL, Kreider BL, Rovera G, Knowles BB (1993) Phenotypically diverse mouse thymic stromal cell lines which induce proliferation and differentiation of hematopoietic cells. Eur J Immunol 23:1201–1214

Fass D, Blacklow S, Kim PS, Berger JM (1997) Molecular basis of familial hypercholesterolaemia from structure of LDL receptor module. Nature 388:691–693

Green EL (1981) Genetics and probability in animal breeding experiments. Oxford University Press, New York

Hong G, Deleersnijder W, Kozak CA, Van Marck E, Tylzanowski O, Merregaert J (1996) Molecular cloning of a highly conserved mouse and human integral membrane protein (Itm1) and genetic mapping to mouse Chromosome 9. Genomics 31:295–300

Imai K (1997) Mouse chromosome 9. Mamm Genome 7:S159–S175

Kim MG, Chen C, Flomerfelt F, Germain RN, Schwartz RH (1998) A subtractive PCR-based cDNA library made from fetal thymic stromal cells. J Immunol Meth, in press

Kim MG, Schuler W, Bosma MJ, Marcu KB (1988) Abnormal recombination of Igh D and J gene segments in transformed pre-B cells of scid mice. J Immunol 141:1341–1347

Kirchgessner CU, Patil CK, Evans JW, Cuomo CA, Fried LM, Carter T, Oettinger MA, Brown JM (1995) DNA-dependent kinase (p350) as a candidate gene for the murine SCID defect. Science 267:1178–1183

Kozak CA, Peyser M, Krall M, Mariano TM, Kumar CS, Pestka S, Mock BA (1990) Molecular genetic markers spanning mouse chromosome 10. Genomics 8:519–524

Lahoud M, Vremec D, Boyd RL, Shortman K (1993) Characterization of thymic nurse-cell lymphocytes using an improved procedure for nurse-cell isolation. Dev Immunol 3:103–112

Lieber MR, Hesse JE, Lewis S, Bosma GC, Rosenberg N, Mizuuchi K, Bosma MJ, Gellert M (1988) The defect in murine severe combined immune deficiency: joining of signal sequences but not coding segments in V(D)J recombination. Cell 55:7–16

Nehls M, Kyewski B, Messerle M, Waldschutz R, Schuddekopf K, Smith AJ, Boehm T (1996) Two genetically separable steps in the differentiation of thymic epithelium. Science 272:886–889

Nehls M, Pfeifer D, Schorpp M, Hedrich H, Boehm T (1994) New member of the winged-helix protein family disrupted in mouse and rat nude mutations. Nature 372:103–107

Ruoslahti E (1996) RGD and other recognition sequences for integrins. Annu Rev Cell Dev Biol 12:697–715

Rawlings ND, Barrett AJ (1994) Families of serine peptidase. In: Barrett AJ (ed) Methods in enzymology, Vol. 244, Academic press, San Diego pp 19–61

Samelson LE, Schwartz RH (1983) T cell-specific alloantisera that inhibit or stimulate antigen-induced T cell activation. J Immunol 131:2645–2650

Shinkai Y, Rathbun G, Lam KP, Oltz EM, Stewart V, Mendelsohn M, Charron J, Datta M, Young F, Stall AM, et al. (1992) RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. Cell 68:855–867

Small M, Kaiser M, Tse W, Heimfeld S, Blumberg S (1996) Activity of neutral endopeptidase and aminopeptidase N in mouse thymic stromal cells which bind double positive thymocytes. Eur J Immunol 26:961–964

Sullivan SL, Adamson MC, Ressler KJ, Kozak CA, Buck LB (1996) The chromosomal distribution of mouse odorant receptor genes. Proc Natl Acad Sci USA 93:884–888

van Ewijk W (1991) T-cell differentiation is influenced by thymic microenvironments. Annu Rev Immunol 9:591–615

Vukmanovic S, Grandea AGD, Faas SJ, Knowles BB, Bevan MJ (1992) Positive selection of T-lymphocytes induced by intrathymic injection of a thymic epithelial cell line. Nature 359:729–732

Vukmanovic S, Jameson SC, Bevan MJ (1994a) A thymic epithelial cell line induces both positive and negative selection in the thymus. Int Immunol 6:239–246

Vukmanovic S, Stella G, King PD, Dyall R, Hogquist KA, Harty JT, Nikolic-Zugic J, Bevan MJ (1994b) A positively selecting thymic epithelial cell line lacks costimulatory activity. J Immunol 152:3814–3823