ORIGINAL ARTICLE

Genome wide mapping of transcriptional start sites in bovine mammary epithelial cells using cap analysis of gene expression (CAGE)

Tomoko Matsubara[1)], Yoshio Kiku[2)], Tomomi Ozawa[1)], Shigeki Inumaru[1)], Hisashi Aso[3)], Takahiro Yamaguchi[3)], David Reynolds[4)], Toshiro Arai[5)] and Tomohito Hayashi[2)]*

1) Pathology and Pathophysiology Research Division, National Institute of Animal Health (NIAH), National Agriculture and Food Research Organization (NARO)
(3-1-5 Kan-nondai, Tsukuba, Ibaraki 305-0856, Japan)
2) Dairy Hygiene Research Division, NIAH, NARO
(4 Hitsujigaoka, Toyohira, Sapporo, Hokkaido 062-0045, Japan)
3) Cellular Biology Laboratory, Graduate School of Agricultural Science, Tohoku University
(1-1 Tsutsumidori-Amamiyamachi, Aoba-ku, Sendai, Miyagi 981-8555, Japan)
4) Genomics Core, Albert Einstein College of Medicine
(1301 Morris Park Avenue, Bronx, New York 10461-1602, USA)
5) Department of Veterinary Science, School of Veterinary Medicine, Nippon Veterinary and Life Science University
(1-7-1 Kyonancho, Musashino, Tokyo 180-8602, Japan)
*To whom correspondence should be addressed: Tomohito Hayashi.
Tel: +81-11-851-2175 Fax: +81-11-853-0767 E-mail: hayatomo@affrc.go.jp

[Abstract]

Mastitis is the most frequent infectious disease in dairy cattle and causes serious economic damage. The identification of a gene regulatory network including signal pathways involved in immune response enabled us to explain a pathogenic mechanism of mastitis. However, studies on bovine mastitis fall behind human and mouse studies, especially expression analysis studies. Recent works revealed that the majority of the genome is transcribed into non-coding RNA (ncRNA) and alternative promoters play important roles in controlling gene expression, thus it is important to profile the bovine transcriptome. Cap analysis of gene expression (CAGE) makes it possible to identify transcriptional start sites (TSS) and quantify the expression levels at each individual TSS. Precise analysis of TSS distribution enables us to predict promoter regions and identify novel genes. It was our goal to obtain unsupervised genome-wide expression data for transcript and promoter identification in the bovine genome. As a model system for this study we used bovine mammary epithelial cells (BMEC), because they are commonly used for studying mastitis. In this study, we show that there were huge amounts of active TSS in BMEC as identified by CAGE. These TSS clusters located in 4,623 intergenic regions

Received: 21 December 2011
Accepted: 2 February 2012

suggested the existence of candidate novel promoters. These results confirm the necessity of global analysis of transcripts to uncover the precise molecular mechanisms of mastitis using BMEC. This study is the first comprehensive analysis of TSS in BMEC by a genome wide scan using CAGE.
Key words: Bovine mammary epithelial cell, cap analysis of gene expression, transcriptional start site, transcriptome.

[Introduction]

Mastitis, the most frequent infectious disease in dairy cattle, has detrimental effects on the quantity and quality of milk production and causes serious economic damage [14]. In current research on mastitis, gene expression analysis has been done to understand inflammation mechanisms and immune response. Recently, studies of bovine genetic research using microarrays led to the construction of a gene network including pathways involved in immune response [8, 11]. The annotation of the bovine genome falls behind the human and mouse genome because the bovine transcriptome has not been studied to the same extent as the human or mouse counterparts. Therefore, currently available cow expression arrays are most likely falling short on covering all genes, and further unsupervised experimental data are needed to improve cow genome annotations. Genomic sequences require annotation based on experimental evidence for their use in data analysis and interpretation.

Contrary to previous notions, recent studies have shown that more than 72% of the mouse genome is transcribed as RNA [12]. Genome scale analysis of the human and mouse transcriptomes using cap analysis of gene expression (CAGE), revealed huge numbers of transcription start sites (TSSs) that are far more abundant than previously expected [4]. There are numerous transcripts derived not only from coding regions but also from non-coding regions. Recent evidence suggests roles of non-coding RNAs (ncRNAs) in regulating transcription [10] and involvement in disease onset [16]. Hence, the transcriptional regulation of ncRNA is the subject of many studies today. Recent studies reveal the transcriptional regulatory network that drives the expression of genes and ncRNA [17]. We can predict that there may be a large number of TSS in bovine mammary epithelial cells (BMEC) similarly to human and mouse cells. For a better understanding of transcriptional regulation, it is important to identify TSS in BMEC.

CAGE makes it possible to identify TSS and quantify the expression levels at each individual TSS. The main step in CAGE is constructing and sequencing DNA tags derived from the initial 27 nucleotides at the 5' ends of mRNAs. By detecting TSSs, we can obtain the transcriptional profile, identify novel genes and predict promoter regions including alternative promoters and binding sites of transcription factors. CAGE-tag based expression measurements provide insight into the regulation of transcription initiation. From CAGE databases, recent studies discovered transcription products (e.g., ncRNA and alternatively spliced transcripts) and the functional domains on the genome, which are involved in transcriptional regulation (e.g., promoter and enhancer). Moreover, analysis of CAGE data eventually allows the inference of transcriptional regulatory networks [5].

Focusing on expression analysis for a better understanding of biological processes in bovine mastitis, it was our goal to obtain unsupervised genome-wide expression data for transcript and promoter identification in the bovine genome. As a model system for this study we used BMEC, because the transcriptome of the cell has not yet been identified although it is a commonly used cell line for

studying mastitis. This study is the first comprehensive analysis of TSS in BMEC. Our results show that TSSs in BMEC localize to several regions including intron and intergenic regions.

[Materials and methods]

Bovine mammary epithelial cells (BMEC)

We used lined BMEC isolated from bovine mammary epithelium. BMEC was established from the mammary gland isolated from a 200-day pregnant Holstein cow [13]. The cells were cultured in Dulbecco's Modified Eagle's Medium (Invitrogen, Carlsbad, U.S.A.), supplemented with 20% inactivated fetal bovine serum, apo-transfferin (10 μg/ml), sodium acetate (5 mM), penicillin (10 IU/ml), and streptomycin (10 μg/ml) at 37°C in air plus 5% CO_2. We seeded 1×10^6 BMEC into 75 cm^2 flask (BD Biosciences, San Jose, USA) and the cells were grown to confluence in culture medium. The cells had typical morphological features: a monolayer, cobblestone, epithelial-like morphology, with close contact between cells. Then, adherent cells were washed with phosphate-buffered saline (PBS) and released using sucrose buffer and 0.025% trypsin (Invitrogen).

CAGE library preparation

Total RNA was extracted from the BMEC using TRIzol (Invitrogen) and used to prepare a CAGE library. The CAGE library was purchased from DNAFORM (Yokohama, Japan), who prepared the CAGE library following the protocol described by Kodzius *et al*., [9] modified by using adaptors suitable for direct sequencing on an Illumina GAII platform. In brief, cDNA complementary strands were synthesized from total RNA by using a mixture of random and oligo-dT primers. The 5' end of cDNA was selected by using the cap-trapper method and cDNA was ligated to a linker containing a recognition site for EcoP15I. After the second strand was synthesized, EcoP15I cleaved cDNAs at a sequence 27 nucleotides away from the 5' end to produce the CAGE tags. Next, a linker was attached to the 3' end of the tag sequence for amplification [Fig. 1(1)].

Sequencing and mapping

Sequencing of the CAGE library was performed on a Genome Analyzer II (Illumina, San Diego, CA, USA) [Fig. 1(2)]. Mapping of CAGE-tag sequences to bovine genome (NCBI Btau 4.0) was done at Genomatix (Genomatix Software, München, Germany). Up to two mismatches (mutations, insertion and/or deletion) were permitted during mapping of the CAGE tags. We used cluster analysis for genome wide identification of local enrichments of CAGE tags representing TSSs obtained from Genomatix (Genomatix Software). Neighboring TSSs with a high probability of expression in a constant proportion were joined into clusters [Fig. 1(3)]. All TSS clusters were correlated with transcripts annotated in the ElDorado database (Genomatix, NCBI Btau 4.0 Version 07-2009) and assigned to these region types: exons, introns or intergenic regions. Regions which included an exon and neighboring intron or intergenic region were categorized as "partial".

[Results]

Definition of TSS by CAGE tags

CAGE tags are 27-nt sequence tags derived from the 5' end of mRNA in the proximity of the cap site. Mapping CAGE tags onto genomic regions identifies TSSs. From sequencing CAGE library, we obtained 3,775,288 CAGE tags from BMEC (Fig. 2). We mapped tags on the bovine genome according to tag sequences, of these CAGE tags, 1,108,153 tags (29.4%) were uniquely hit to one position in the genome (unique-hit), 253,192 tags (6.7%) were mapped to more than two regions on the genome [207,938 tags (5.5%) were less than 50 regions (multi-hit) and 45,254 tags (1.2%) were more than 50 regions (ambiguous)], 2,413,910 tags (63.9%) were insufficiently mapped and 33

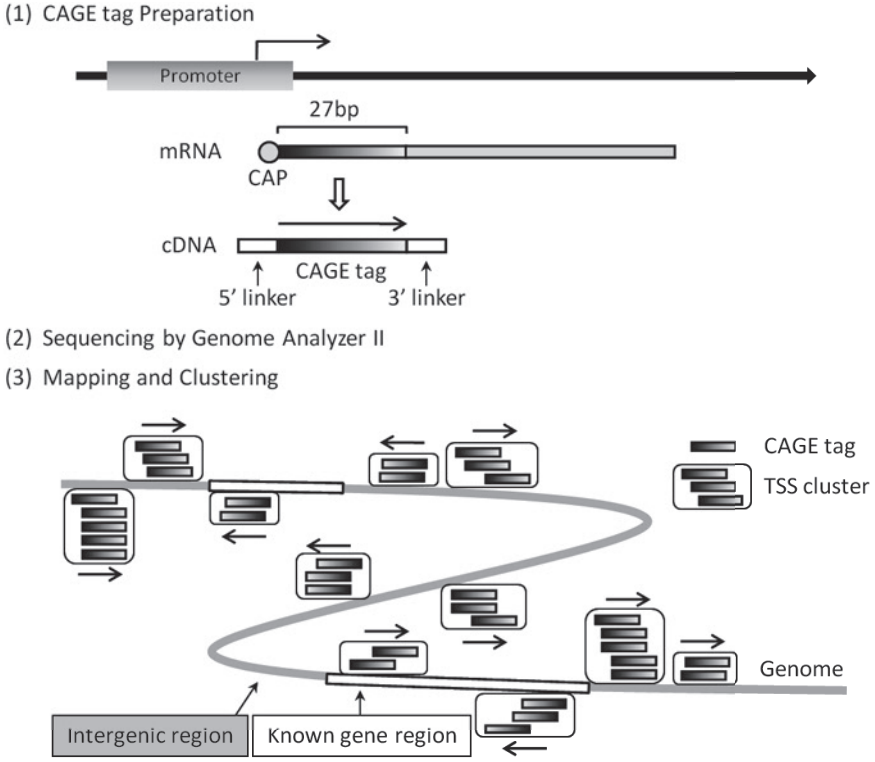

Fig. 1 CAGE tag preparation, mapping to bovine genome and distribution by region of CAGE tags. (1) CAGE tags are 27bp sequence tags derived from the mRNA sequenced in the proximity of the cap site. 5' linker and 3' linker were ligated to CAGE tags. (2) CAGE library was sequenced on a Genome Analyzer II. (3) CAGE tags were mapped to the bovine genome and nearby CAGE tags were clustered into a single TSS cluster.

tags (0.0%) were ignored.

Identifying promoters by CAGE tags (unique hit tags)

We clustered nearby CAGE tags into TSS cluster. The TSS clusters mapped not only to known promoter sites but also unconventional sites such as introns and intergenic regions. Of the total 19,197 regions which TSS clusters mapped into, 6,949 regions (36.2%) map in exon sites, 5,766 regions (30.0%) partially overlap an exon and an intron/intergenic region, 1,859 regions (9.7%) map in intron sites and 4,623 regions (24.1%) map in intergenic regions (Fig. 3). Based on the bovine genome annotation, 6,681 CAGE regions (34.8%) were annotated as known promoters.

[Discussion]

ncRNA has been implicated in control of genetic networks. Many studies demonstrated the existence of tissue-specific transcripts [7] and the genetic diversity within breeds [1]. Alternative splicing has been noted as a factor of the diversity of gene expression [2]. Considering these things, it is necessary to examine TSS for better understanding of transcriptional regulation. Thus, we identified complete set of TSSs in BMEC.

Recent studies revealed that large proportions of the mouse and human genome are transcribed into RNA [4]. In this study, similarly, we show that there are large amounts of CAGE tag clusters in bovine genome. Furthermore, we found TSS not only in already-known promoter regions but also in intergenic regions. Thus, it was revealed that there were 4623 regions as candidate novel transcripts and suggested the existence of novel promoters.

It was reported that ncRNA is a major component of the transcriptome in mouse [3]. Recent work suggests that ncRNA can play critical roles in a wide range of cellular processes,

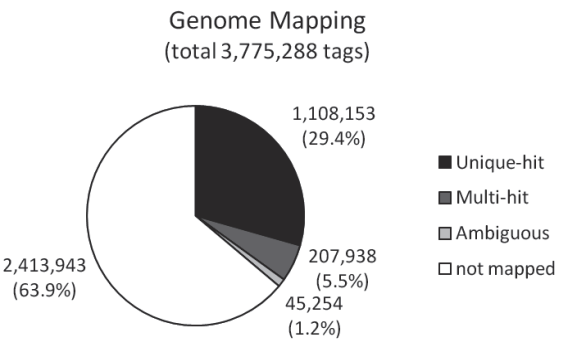

Fig. 2 Mapping CAGE tags to bovine genome. The upper value is the absolute number of tags and the bottom value is the percentage of the type of tag to total tags.

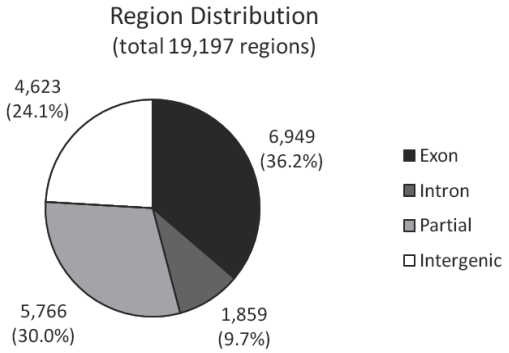

Fig. 3 Region distribution of TSS clusters to genomic elements. The upper value is the number of regions overlapped with genomic elements and the bottom value is the percentage of the type of region to total regions.

from protein secretion to gene regulation [15]. Furthermore, because ncRNA is considered to be related to disease onset [16], understanding ncRNA is very important in the development of therapy and drug. However, in the bovine genome, ncRNA have scarcely been studied. In this study, there was a multitude of TSSs mapped to introns and intergenic regions, which are possibly ncRNAs. Our data will be useful for identifying ncRNAs in the cow genome.

In mouse, the CAGE tags mapped to single regions were approximately 61.8%, multi-region were 14.4% and the other 23.8% of tags could not be mapped [6]. Contrary to these reports, this study found that uniquely mapped tags were only 29.3% of the CAGE tags, multiple mapped tags were 5.5% and 64.9% were not mapped in this study. It is possible that this low efficiency of mapping to genome is attributed to insufficiency of the available bovine genome and diversity within breeds. Similarly, genetic functions in cattle are poorly understood. In future, with progress in genetic analysis and genome annotation of Holstein breed, this data obtained from the CAGE tags can provide more details on the bovine transcriptome.

In this study, we obtained unsupervised genome-wide expression data for genes and their promoters in the bovine genome. We have shown that there was a huge amount of TSS in BMEC. This TSS data potentially indicates the existence of novel transcripts and promoters. Further studies are needed to identify promoters from these TSSs. This data set will be useful for understanding the diversity within breeds and epigenetic control of transcription. With advances in annotation of bovine genome, this data set will allow us to better understand the biological processes in bovine mastitis.

[Acknowledgements]

This work was supported in part by a grant-in-Aid from mastitis treatment project of the Ministry of Agriculture, Forestry and Fisheries of Japan and the Strategic Research Base Development Program for Private Universities from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), 2008-2012.

[References]

1. Bovine HapMap Consortium, 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. 324, 528-532.
2. Cáceres, J. F., Kornblihtt, A. R., 2002. Alternative splicing: multiple control mechanisms and involvement in human disease. TRENDS in Genetics. 18, 186-193.
3. FANTOM Consortium; RIKEN Genome Exploration Research Group Phase I & II Team,

2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 420, 563-573.

4. FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). 2005. The transcriptional landscape of the mammalian genome. Science. 309, 1559-1563.
5. FANTOM Consortium, 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat. Genet. 41(5), 553-562.
6. Faulkner, G. J., Forrest, A. R., Chalk, A. M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D. A., Grimmond, S. M.. 2008. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics. 91(3), 281-288.
7. Fürbass, R., Kalbe, C., Vanselow, J., 1997. Tissue-specific expression of the bovine aromatase-encoding gene uses multiple transcriptional start sites and alternative first exons. Endcrinol. 138, 2813-2819.
8. Günther, J., Esch, K., Poschadel, N., Petzl, W., Zerbe, H., Mitterhuemer, S., Blum, H., Seyfert, H. M., 2011. Comparative kinetics of Escherichia coli- and Staphylococcus aureus-specific activation of key immune pathways in mammary epithelial cells demonstrates that S. aureus elicits a delayed response dominated by interleukin-6 (IL-6) but not by IL-1A or tumor necrosis factor alpha. Infect. Immun. 79(2). 695-707.
9. Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., Carninci, P., 2006. CAGE: cap analysis of gene expression. Nat. Methods. 3(3), 211-222.
10. Mattick, J. S., 2009. The genetic signatures of noncoding RNAs. PLoS. Genet. 5(4), e1000459.
11. Moyes, K. M., Drackley, J. K., Morin, D. E., Bionaz, M., Rodriguez-Zas, S. L., Everts, R. E., Lewin, H. A., Loor, J. J., 2009. Gene network and pathway analysis of bovine mammary tissue challenged with Streptococcus uberis reveals induction of cell proliferation and inhibition of PPARγ signaling as potential mechanism for the negative relationships between immune response and lipid metabolism. BMC Genomics. 10: 542.
12. RIKEN Genome Exploration Research Group, Genome Science Group, the FANTOM Consortium, 2005. Antisense transcription in the mammalian transcriptome. Science. 309, 1564-1566.
13. Rose, M. T., Aso, H., Yonekura, S., Komatsu, T., Hagino, A., Ozutsumi, K., Obara, Y., 2002. In vitro differentiation of a cloned bovine mammary epithelial cell. J. Dairy Res. 69, 345-355.
14. Seegers, H., Fourichon, C., Beaudeau, F., 2003. Production effects related to mastitis and mastitis economics in dairy cattle herds. Vet. Res. 34, 475-491.
15. Szymański, M., Barciszewska, M. Z., Zywicki, M., Barciszewski, J., 2003. Noncoding RNA transcripts. J. Appl. Genet. 44(1), 1-19.
16. Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., Mattick, J. S., 2010. Non-coding RNAs: regulators of disease. J. Pathol. 220, 126-139.
17. Turner, A. M. and Morris, K. V., 2010. Controlling transcription with noncoding RNAs in mammalian cells. Biotechniques. 48(6), ix-xvi.

ウシ乳腺上皮細胞における CAGE 解析法を用いた ゲノム全域にわたる転写開始点の解析

松原朋子[1)]　菊　佳男[2)]　尾澤知美[1)]　犬丸茂樹[1)]　麻生　久[3)]
山口高弘[3)]　David Reynolds[4)]　新井敏郎[5)]　林　智人[2)*]
1）動物衛生研究所　病態研究領域（〒305-0856　茨城県つくば市観音台 3-1-5）
2）動物衛生研究所　寒地酪農衛生研究領域（〒062-0015　北海道札幌市豊平区羊ヶ丘 4）
3）東北大学　大学院　農学研究科　動物機能科学講座（〒981-8555　宮城県仙台市青葉区堤通雨宮町 1-1）
4）Genomics Core, Albert Einstein College of Medicine（1301 Morris Park Avenue, Bronx, New York 10461-1602, USA）
5）日本獣医生命科学大学　獣医学部　獣医生理化学教室（〒180-8602　東京都武蔵野市境南町 1-7-1）
＊連絡者：林　智人
Tel：011-851-2175　Fax：011-853-0767　E-mail：hayatomo@affrc.go.jp

［要　約］

乳房炎は病原微生物の乳房内感染によって引き起こされ、乳質ならびに泌乳量の低下を招く疾病である。その経済的損失は大きく、酪農経営を脅かす問題である。免疫応答に関わる遺伝子制御機構や細胞内シグナル伝達経路が同定されることによって、近年様々な病気の発症機序の解明が進みつつある。しかし、乳房炎研究におけるアプローチは遺伝子研究の観点からすると遅れを取っている。一方、ノンコーディング RNA（ncRNA）や選択的プロモーターが遺伝子発現の調節に重要な役割を果たすことが明らかになってきており、転写産物の網羅的解析が必要になってきている。遺伝子発現プロファイルやプロモーターを特定する新しいアプローチ方法として最近開発された Cap Analysis Gene Expression（CAGE）法は、転写開始点（TSS）の位置と発現量を解析することができ、TSS の正確な分布情報はプロモーター領域の予測や新規遺伝子の発見を可能にしている。本研究では乳房炎発症に関与するウシゲノムの転写産物とプロモーターの同定のためのゲノム全域にわたる発現データを得ることを目的とし、乳房炎の研究に一般的に用いられているウシ乳腺上皮細胞（BMEC）を用いた CAGE 解析を行った。その結果、膨大な数の転写開始点がゲノム上に特定され、これらの TSS クラスターが 4623 の遺伝子間領域に位置していることが明らかになり、さらに新規プロモーターが存在することが示唆された。本研究は CAGE 解析により BMEC の TSS を網羅的に解析した初めての研究である。今後さらに BMEC を用いて乳房炎の発症機序に関与する遺伝子動態を理解するためには、個々の転写産物におけるより詳細な解析が必要になると考えている。

［方　法］

CAGE 解析の概略：①全 RNA から相補鎖 cDNA の合成、②キャップトラッピング法による 5’ 末端の選別、③ EcoP151 を含むビオチン化したリンカーの結合、④第二鎖 cDNA の合成、⑤ 5’ 末端から 27 塩基の切り出しの後に、⑥この配列の 3’ 側へのリンカーの結合を行い、⑦ビオチン化した cDNA タグ（CAGE タグ）のみ抽出して精製した［Fig. 1（1）］。得られた CAGE タグの配列をシークエンシングにより決定し［Fig. 1（2）］、ゲノム上にマップした［Fig. 1（3）］。

［結　果］

BMEC から全部で 3,775,288 の CAGE タグが得られた。そのうち、29.4％はゲノム上の 1 ヶ所（ユニークヒット）、6.7％は 2 ヶ所以上（マルチヒット）にマッピングされ、63.9％はマッピングできなかった（Fig. 2）。また、ゲノム上で隣接する CAGE タグを TSS クラスターとしてクラスター化したところ、クラスターがマッピングされたゲノム上の領域は 19197 ヶ所あった。そのうち、36.2％はエクソン、9.7％はイントロン、30.0％はエクソン－イントロン／遺伝子間領域に跨いだ領域、24.1％は遺伝子間領域であった。ウシゲノムのアノテーションに基づくと、34.8％の領域がプロモーターとして知られている領域であった。

[考　察]

今回、BMEC において膨大な数の TSS が見付かった。多数の TSS がイントロンや遺伝子間領域にマッピングされたことから、本解析により、新規プロモーターや ncRNA の同定に役立つデータが得られたと考えている。マウスに比べて、マッピング効率が低かったが、その原因として、ウシゲノムの情報が不十分なことや品種間の違いが考えられる。今後、ウシゲノムのアノテーションが進むのにつれ、このデータセットはより詳細な生物学的プロセスを理解するために有用な情報源となりうると考えている。以上の結果から、BMEC を用いて乳房炎の発症機序に関与する遺伝子動態を理解するためには、個々の転写産物における詳細な解析が必要となると考えられる。