Research

Dynamic usage of transcription start sites within core promoters

Hideya Kawaji[*], Martin C Frith[†‡], Shintaro Katayama[†], Albin Sandelin[†§], Chikatoshi Kai[†], Jun Kawai[§¶], Piero Carninci[§¶] and Yoshihide Hayashizaki[†¶]

Addresses: [*]NTT Software Corporation, 209 Yamashita-cho Nakak-ku, Yokohama, Kanagawa, 231-8551, Japan. [†]Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan. [‡]Institute for Molecular Bioscience, University of Queensland, 306 Carmody Road, Brisbane, Queensland 4072, Australia. [§]The Bioinformatics Centre, University of Copenhagen, Universitetsparken 15, DK-2100 København Ø, Denmark. [¶]Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan.

Correspondence: Hideya Kawaji. Email: kawaji@po.ntts.co.jp. Shintaro Katayama. Email: rgscerg@gsc.riken.jp

Published: 12 December 2006

Genome ***Biology*** 2006, **7:**R118 (doi:10.1186/gb-2006-7-12-r118)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/12/R118

Received: 31 July 2006
Revised: 26 October 2006
Accepted: 12 December 2006

Abstract

Background: Mammalian promoters do not initiate transcription at single, well defined base pairs, but rather at multiple, alternative start sites spread across a region. We previously characterized the static structures of transcription start site usage within promoters at the base pair level, based on large-scale sequencing of transcript 5' ends.

Results: In the present study we begin to explore the internal dynamics of mammalian promoters, and demonstrate that start site selection within many mouse core promoters varies among tissues. We also show that this dynamic usage of start sites is associated with CpG islands, broad and multimodal promoter structures, and imprinting.

Conclusion: Our results reveal a new level of biologic complexity within promoters - fine-scale regulation of transcription starting events at the base pair level. These events are likely to be related to epigenetic transcriptional regulation.

Background

There is great interest in elucidating the control of transcription initiation, because these controls are major components of the gene regulatory networks that underlie the development and diversity of animals [1,2]. The standard view is that regulatory action takes place at distal and proximal enhancer and repressor *cis* elements, which are bound by transcription factors that interact with the basal transcription machinery at the core promoter to influence transcription. In this view, core promoters themselves are functionally simple, but recent data reveal that they are structurally complex, with a range of alternative transcription start sites (TSSs) at the base pair level [3-5]. A key issue is whether these complex structures are just 'biologic noise' from imprecise binding of basal transcription factors or whether TSS selection is precisely regulated.

Cap analysis of gene expression (CAGE) is a method used to identify TSSs and, at the same time, to measure their expression levels by counting a large number of sequenced 5' ends of full-length cDNAs, termed CAGE tags [6,7]. The advantage of this method is that it provides a view at base pair level of the expression profiles of TSSs even within a promoter. In contrast, the most commonly used high-throughput

methodology for measuring gene expression, namely the microarray, profiles transcript expression without distinguishing between alternate 5' ends. Expressed sequence tag (EST) and full-length cDNA sequencing characterize end structures of transcripts, but their quantification ability is limited because of their cost. Additionally, some cDNA libraries are subtracted or normalized for exploration of novel transcripts, and these libraries cannot provide a quantitative view of expression [8,9].

In the FANTOM3 (functional annotation of mouse 3) project, the CAGE method was applied to more than 20 tissues from mouse and human [4,10]. More than seven million mouse CAGE tags were sequenced and mapped to the mouse genome, and so many core promoters are represented by many CAGE tags. This gives unprecedented opportunities to resolve the internal structures of core promoters.

As with cDNA sequencing, sequencing a large number of CAGE tags may capture errors, such as degraded transcripts or incomplete cDNA synthesis events. Extensive experimental and statistical validation of the CAGE set analyzed in this study, presented elsewhere (see the report by Carninci and coworkers [4] and its supplementary material), demonstrated good reliability even for single CAGE tags. A potential weakness with the method is the tag length (20-21 base pairs [bp]); with only a few sequencing errors, mapping tags back to the genome can be problematic. In the present study we used only unequivocal tag mappings [4] and focused on core promoters with more than 100 co-occurring tags. Another general issue with all tag-based technology is how to reliably associate tags with their corresponding full-length transcript; however, this is not a CAGE-specific problem and similar challenges are faced when using array-based methods.

Interestingly, transcription initiation was found to occur at multiple nucleotide positions within a core promoter region in many cases, although the start sites are more tightly clustered (but still not uniquely defined) for a subset of promoters with an over-representation of TATA boxes. Thereby, most core promoters do not have a single TSS but rather an array of closely located initiation sites. For clarity, this is conceptually different from alternative promoters, in which core promoters are separated by clear genomic space. In order to analyze arrays of tags corresponding to core promoters it is necessary to cluster adjacent tags [10]. A tag cluster is defined as a segment of a chromosome, on either the forward or reverse strand, where each 20 bp subregion contains at least one transcript 5' end identified by RIKEN full-length cDNAs, RIKEN-5' ESTs [10], GIS ditags [11], GSC ditags [12], or CAGE tags [7].

We previously found that the TSS distributions of tag clusters have various 'shapes'. This means that there are various modes in selection of transcription initiation sites depending on promoters. In our previous study, tag clusters with sufficient (100 or more) CAGE tags for statistical analysis (1.1% [8,157] of the 736,403 tag clusters) were classified into four shape classes (for representative examples, see Figure 1): a single dominant peak (1,875 tag clusters), a general broad distribution (2,702), a broad distribution with a dominant peak (1,880), and a bimodal or multimodal distribution (1,700). Only the first class (23% of the 8,157) represents a narrowly defined TSS location, whereas the remaining classes are categories of broad regions with multiple TSSs. The single dominant peak class is associated with TATA boxes and tissue-specific expression, and the broad classes are associated with CpG islands and ubiquitous expression [4,10]. Although a classical model of transcriptional regulation can account for the single dominant peak class, it cannot explain arrays of TSS and their lack of TATA boxes. Because the shapes generally are very similar between human and mouse orthologous promoter regions, these properties strongly suggest that different modes of TSS selection exist between different promoter types [4].

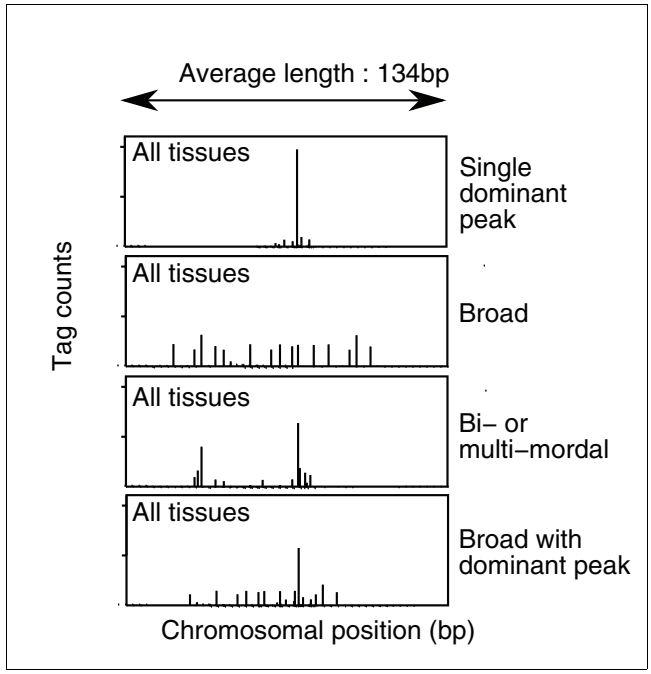

Figure 1
Four shape classes of static TSS usage. Tag clusters were classified into four classes based on CAGE tag counts from all tissues. The tag counts are displayed by histograms, where the x-axis indicates genomic coordinates or chromosomal location, and the y-axis indicates the total counts of CAGE tags. bp, base pairs; CAGE, cap analysis of gene expression; TSS, transcription start site.

A basic issue that must be addressed if we are to understand such broad transcription start regions is whether start site selection is precisely regulated or whether TSS usage is driven by nonspecific binding of basal transcription factors [13]. If TSS selection is regulated, then broad start regions could be caused by varying concentrations of transcription factors that

favor initiation at different sites [14] or by epigenetic mechanisms such as DNA methylation, histone modifications, and chromatin remodeling [15-20]. If this is true, then it would be possible for the cell to modify the start site selection within a promoter in different contexts (such as tissues). On the other hand, if start site selection is primarily driven by the properties of the genomic sequence, then we would not expect major differences in TSS selection between tissues in a given broad promoter.

To address this issue, we examine tissue specificity at the base pair level, or fine-grained tissue-specific usage of TSSs. Note that our focus is not on alternative promoters, which are multiple promoters used by the same gene [4,21]. Rather, we investigate alternative TSSs within a core promoter region.

Here, we show that there are distinct, tissue-specific modes of start site selection within core promoters. To suggest possible mechanisms for this phenomenon, we show that such fine-grained tissue specificities of TSSs are associated with some expression contexts, such as tag cluster shapes, and genomic imprinting candidates.

Results and discussion

Tested tag clusters

We will be able to identify reliably only large usage biases if a tag cluster has few tags from each tissue, whereas more subtle biases will be reliably detectable if a tag cluster has many tags from some tissues. From this viewpoint, we use 8,157 tag clusters with 100 or more CAGE tags for statistical analysis. These clusters have previously been classified into the four shape classes based on CAGE tag distributions [4]. The mean length of these tag clusters is 134.2 bp, and 95% of them are under 250 bp in length. The mean lengths of the four classes based on their shapes or CAGE tag distributions are as follows: 87.0 bp for the single dominant peak, 146.7 bp for the broad distribution with a dominant peak, 180.5 bp for the multimodal distribution, and 129.1 bp for the general broad distribution. The mean length for the multimodal class is the longest among the four classes, being over twice the mean length for the single dominant peak.

CAGE tags in a tag cluster come from several tissues, and their accumulation by each tissue and each genomic position is required to uncover dynamic usages of TSSs within a promoter. Figure 2 shows some possible cases of TSS selection within a promoter by different tissues, where panel a is a case of no differences between tissues, and panels b and c show cases of clear tissue specificity. Below, we examine whether the tag clusters have any tissue specificities, based on CAGE tag counts.

Positionally biased promoters

In our exploration of tissue specificities within a tag cluster in which transcripts are initiated over a continuous region, we have no clear border to distinguish subregions to be compared with each other. The situation is different from exploration of alternative promoters, where each promoter is clearly separated by a certain genomic space. To cope with this issue, we adopt two strategies to explore fine-grained tissue specificity as comprehensively as possible: first, we explore differences in central (or median) TSS position depending on tissue; and second, we explore subregions whose expression profiles are different from the rest of the tag cluster. The first strategy can identify an intuitive type of fine-grained tissue specificity, namely overall bias of centered position, such as shown Figure 2b. There remain other types of tissue specificity, such as shows in Figure 2c, which has some internal regions with distinct tissue specificities but no clear differences in terms of the centered position. The second strategy was devised to find these cases.

First, we examined whether the median location of transcription initiation within each tag cluster varies between tissues (Figure 3). This entails subdividing the tag cluster into multiple tag distributions depending on tissue, and then assessing whether the centers of all such tag distributions are similarly positioned. Because of the tag cluster definition, we would expect that some, if not all, of such subdistributions will overlap to some extent with each other, because if a group of tags does not overlap with any other then it would not be part of the initial cluster but would form a distinct alternative promoter. We did not attempt to fit the subdistributions to any generic template such as normal distributions, because the shapes can vary greatly [4] and in some cases there were too few tags to fit the subdistributions. Moreover, at the base pair level start site selection is biased toward pyrimidine-purine dinucleotides (where the transcript starts at the pyrimidine) [4], which makes any normality assumption unsound.

Given the above, we employed a statistical test with no in-built assumption about distributions, namely the Kruskal-Wallis one-way analysis of variance by ranks. It tests the null hypothesis that several samples come from populations with the same median [22] (this is essentially a nonparametric variant of the classical analysis of variance test). Thus, rejection of the null hypothesis implies that at least one of the underlying tag distributions has a distinct center point. The null hypothesis was rejected ($P < 0.01$) for 2,491 out of 8,157 tag clusters (30%), and we term these cases 'positionally biased'. The test does not indicate which tissues differ in median, just that they are not all the same.

An example of a positionally biased tag cluster is shown in Figure 4a. A tag cluster located at the 5' end of *PPap2b* (phosphatidic acid phosphatase type 2B) has two peaks of CAGE tags about 20 bp apart. The downstream peak is the most used and corresponds to the median in liver libraries, whereas the upstream peak is the most utilized in lung. These two regions are clearly utilized in a tissue-specific manner, and this results in a statistically significant difference in

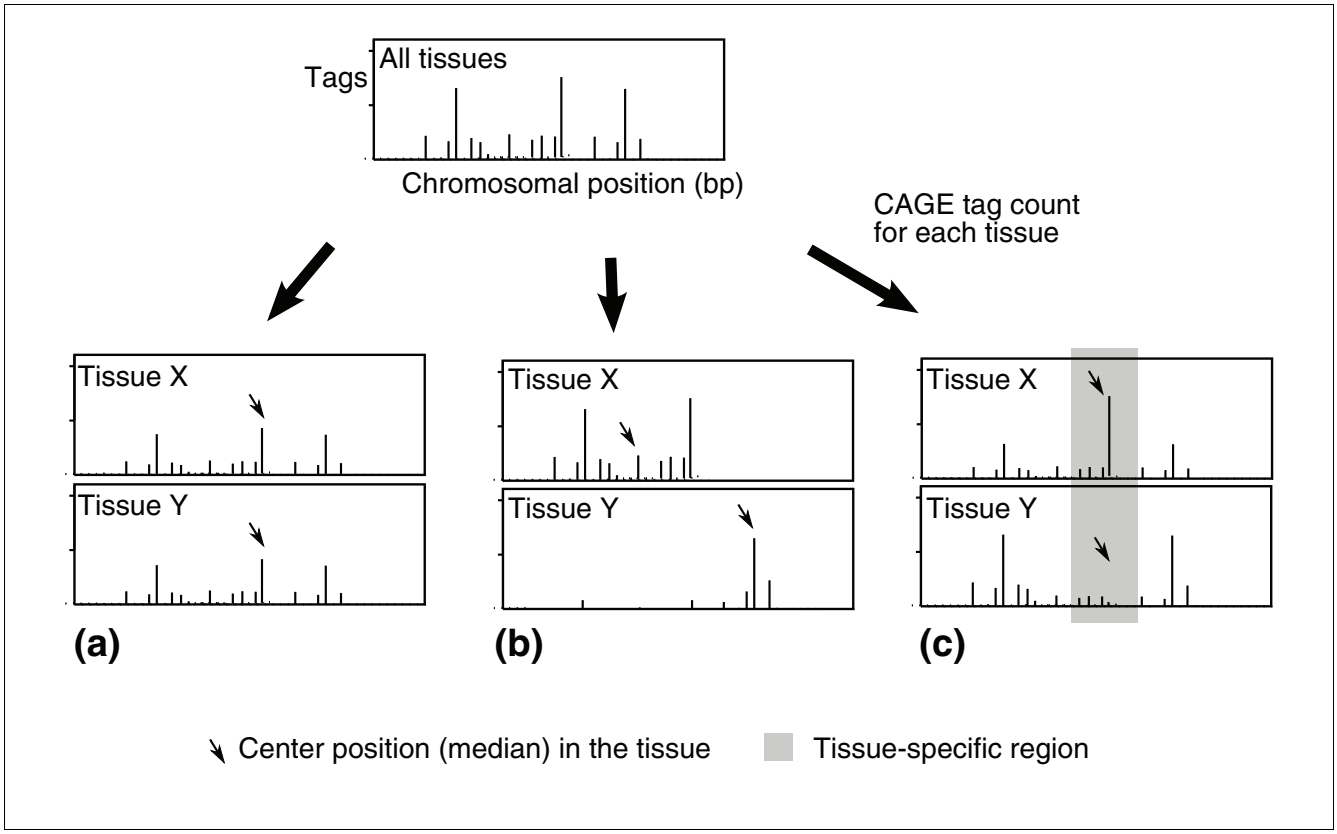

Figure 2
Possible cases of TSS usage among tissues. Possible cases of TSS usage sharing the same static structure of TSS: **(a)** TSS usage is identical between tissues X and Y; **(b)** upstream sites are favored in tissue X whereas downstream sites are favored in tissue Y; and **(c)** some subregions exhibit distinct TSS usage between tissues. The CAGE tag count of each tissue at each position is displayed as a vertical line, where the x-axis indicates genomic coordinates or chromosomal location and the y-axis indicates the CAGE tag count. bp, base pairs; CAGE, cap analysis of gene expression; TSS, transcription start site.

median TSS location. If TSS selection is influenced by distinct but proximal *cis* elements depending on tissues, then this type of TSS usage would be expected.

Regionally biased promoters

Second, we identified tissue-specific subregions of 21 bp within tag clusters, using a Bayesian statistics based method developed previously for analysis of alternative splicing (see Materials and methods, below) [23].

Of the total 8,157 tag clusters, 3,542 (43%) had at least one tissue-specific subregion. As expected, most of the positionally biased clusters (1,541/2,491 [62%]) also had tissue-specific subregions (Figure 5). In total, about half (4,492/8,157 [55%]) of the tag clusters examined exhibit internal tissue-specificity of some kind. Because the positionally biased clusters were already shown to have a tissue bias in TSS selection, we focused on those tag clusters that were not positionally biased but still had subregions with distinct tissue usage. We term these cases, which cover 2,001 out of 8,157 tag clusters (25%), 'regionally biased' (Figure 3).

An example of a regionally biased cluster is shown in Figure 4b. A tag cluster located at the 5' end of *ORF61*, which encodes a 574 amino acid protein of unknown function, has a broad shape, and the median TSS locations are positioned roughly in the center of the tag cluster. Although there is no significant difference of medians among tissues, the CAGE tag distributions in its subregions are different from each other depending on tissues. For example, upstream TSSs are used frequently in embryo whereas downstream TSSs are used frequently in liver. Tissue specificities change along the genome, but the other TSSs in the intermediate region and at both ends contribute to no significant difference in central TSS position.

Associations with CpG islands and CAGE tag shape classes

To explore the context of promoters with dynamic TSS usage, we examined their relations with CpG islands. Of the 5,607 tag clusters located in CpG islands, 1,908 (34%) and 1,650 (29%) are classified as positionally and regionally biased, respectively. Table 1 shows associations between CpG islands,

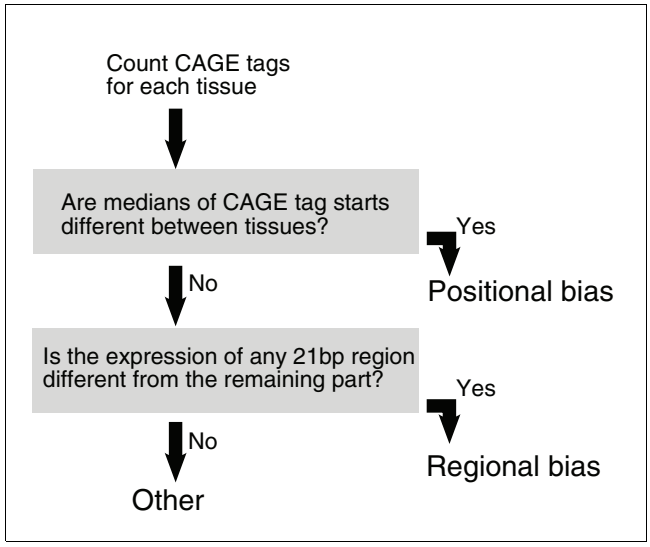

Figure 3
Classification of dynamic TSS usage within promoters. The classification flow of tag clusters. All of the examined tag clusters are classified into three categories - positional bias, regional bias, and others - based on CAGE tag counts from each tissue. bp, base pairs; CAGE, cap analysis of gene expression; TSS, transcription start site.

and positionally and regionally biased promoters. Each cell indicates a one-sided *P* value of the Fisher's exact test for the null hypothesis that the two categories do not have any positive association. For example, the cell in the first row and the first column indicates the result of the statistical test based on a 2 × 2 contingency table, whose columns represent positionally biased and other (not positionally biased) promoters and whose rows represent CpG and other (non-CpG) promoters. Table 1 indicates that both positionally and regionally biased tag clusters are associated with CpG islands with statistical significance ($P < 1.0 \times 10^{-3}$). Tag clusters containing internal regions with different tissue-specificities tend not to be in the single dominant peak class in which transcription starts from a narrowly fixed position. This is to some degree expected just because of the nature of the single dominant peak class, because the width of such promoters is small. These associations are consistent with the previous finding that broad tag clusters are associated with CpG islands [4].

We also examined their relations with shapes of CAGE tag distributions (Table 1). A significant association of positional bias with the multimodal shape class suggests that the multiple peaks are superimposed prominent TSSs utilized in a tissue-specific manner, implying that tag clusters with multimodal shapes consist of multiple and overlapping promoters. This can be expected from the definition of tag clusters, where two proximal and distinct promoters are joined if rarely used TSSs are located between them. Interestingly, Table 1 also shows a significant association of the regionally biased class with the general broad tag distribution. This reveals distinct tendencies between positional and regional biases, and that tag clusters without remarkable peaks are also regulated tissue specifically on a fine-grained scale. Non-specific DNA binding of transcription factors [13] is unlikely to explain these tag clusters.

Associations with imprinting

Genomic imprinting is epigenetic modification of genes whose expression is determined according to their parent of origin [24]. The key molecular mechanism is DNA methylation, which can repress transcription by direct and indirect mechanisms, such as inhibiting the binding of specific transcription factors, and recruiting methyl-CpG-binding proteins associated with repressive chromatin remodeling [25]. Interestingly, different machineries for maternal and paternal silencing have been suggested: maternal repression is effected by promoter methylation of a target transcript, and paternal repression by inactivation of its antisense transcript by maternal methylation [26]. Analysis of *Eed* mutant mice suggests that paternally and maternally inherited chromosomes can use different chromatin silencing mechanisms [27,28]; however, the details remain unclear.

To explore links between dynamic TSS usage and imprinting, we used candidate imprinted transcripts stored in the EICO database [29], which were identified by differential expression dependent upon chromosomal parent of origin using cDNA microarrays [30]. The sensitivity of the method was demonstrated by identification of previously reported imprinted genes [30]. It should be emphasized that the EICO database lists candidate imprinted transcripts and non-imprinted transcripts under the control of imprinted transcripts by identification of differential expression between parthenogenotes and androgenotes [30,31].

We found that 328 of the 8,157 tag clusters used in this study are located at 5' ends of the imprinting candidates, and 115 (35%) and 104 (31%) of them are classified as positionally and regionally biased, respectively. Table 1 shows the statistical significances of their associations with these candidates, which indicates that paternally and maternally imprinted transcripts are associated with positional and regional biases. We also found that paternal and maternal imprinting candidates are associated with the general broad shape class with *P* values of 0.04 and 1.6×10^{-5}, where Fisher's exact test is used for the null hypothesis that paternal imprinting (or maternal imprinting) and the general broad shape class do not have any positive association. It is surprising that paternally imprinted promoters with positional bias are not associated with the multimodal shape class, which is a characteristic of positional bias in general. Although these paternally imprinted promoters are just special cases of positional bias, maternally imprinted promoters may be more representative cases of regional bias.

As an example, *Snrpn*, which encodes small nuclear ribonucleoprotein N, is an imprinted gene related to Prader-Willi

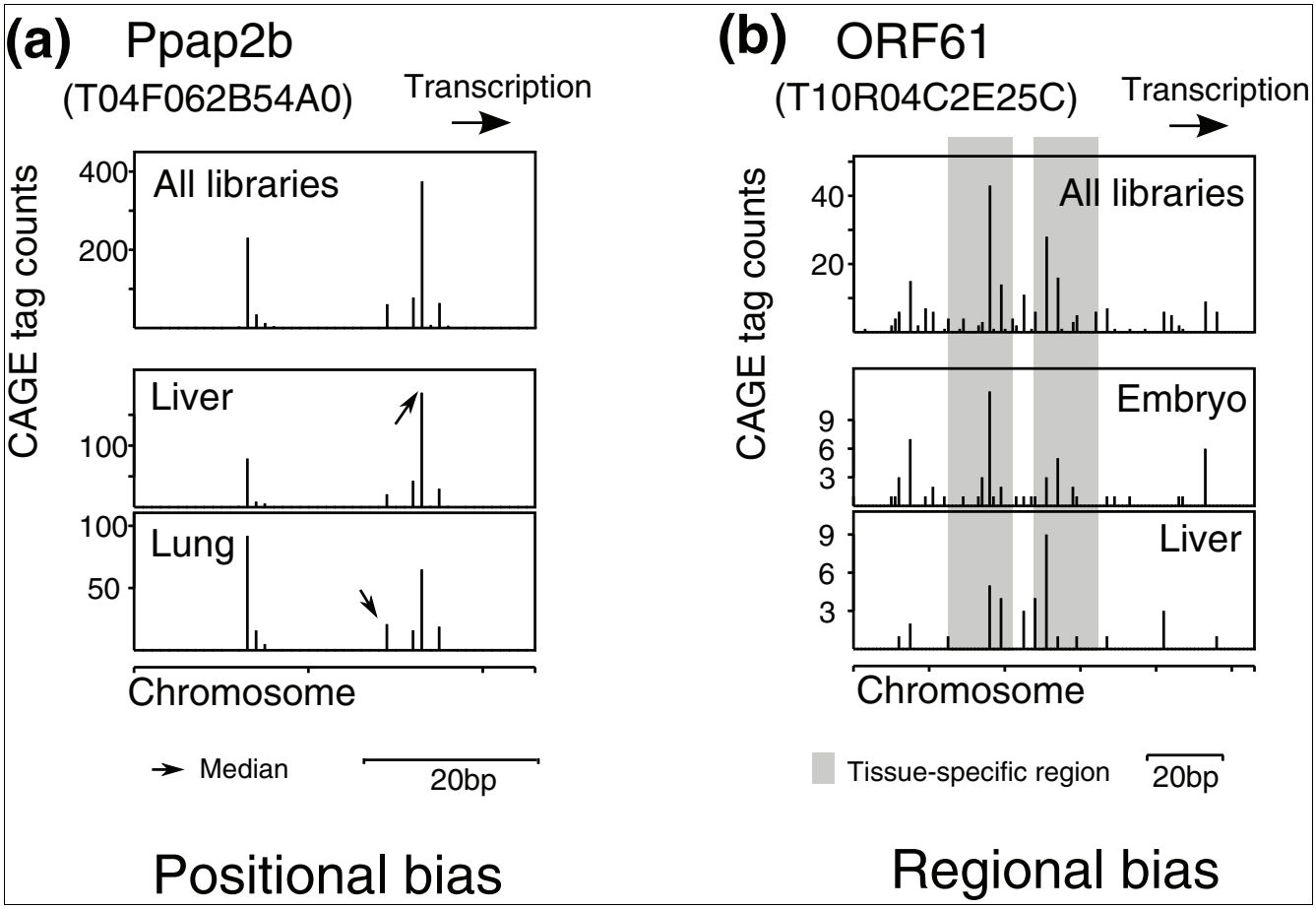

Figure 4
Examples of fine-grained tissue specificity. Examples of fine-grained tissue specificity are shown. **(a)** CAGE tag distribution of a tag cluster located at the 5' end of *Ppap2b* (tag cluster ID: T04F062B54A0), which is classified as positionally biased. The CAGE tag count at each nucleotide position is displayed as a vertical line. Totals of CAGE tag counts derived from all libraries are shown as all libraries, and two tissues picked from the 22 utilized tissues, liver, and lung are also shown. Arrows within the histograms indicate median locations of transcription initiation. **(b)** CAGE tag distribution of a tag cluster located at the 5' end of *ORF61* (tag cluster ID: T10R04C2E25C), which is classified as regionally biased. The CAGE tag counts of all libraries, embryo, and liver are shown. Gray backgrounds indicate regions with different tissue specificity from the remaining part of the tag cluster. CAGE tag counts in each library and additional information for tag clusters is also accessible from the CAGE Analysis Database [39]. CAGE, cap analysis of gene expression.

syndrome, and its 5' end is maternally methylated [32]. The tag cluster T07R02CED41C is located at the 5' end of *Snrpn* and classified as regionally biased. Figure 6 shows the expression profile at the base pair level. Different tissue specificities can be observed in the regions with grey background. As seen in Figure 6, the B region, which exhibits high expression in general, is less expressed in somatosensory cortex and visual cortex, whereas the A region is less expressed in whole brain and somatosensory cortex. Methylation-sensitive polymerase chain reaction (PCR) has revealed that each of the CpG dinucleotides at the 5' end is methylated at different levels in the embryo, especially at 10.5 days *post coitum*, and also revealed that methylation levels change dynamically in a developmental process [33]. An interpretation of this fine-grained tissue specificity is that the differential methylation of each CpG dinucleotide affects the transcription machinery, and results in different specificities without a clear positional bias. This interpretation is based on the fact that specific paternal and maternal methylation of imprinted genes starts at different genomic locations and that CpG methylation gradients may influence transcription. This would affect fine-grained transcription start usage in a 'regionally biased' way among tissues for maternally imprinted genes.

Associations with tissue-specific differentially methylated regions

Methylation is involved in tissue-specific expression in some cases, as well as genomic imprinting. Genome-wide analysis of DNA methylation status using restriction landmark genomic scanning (RLGS) [34] identified chromosomal regions that are differentially methylated in a tissue-specific manner [35,36]. Quantitative real-time PCR and bisulfite

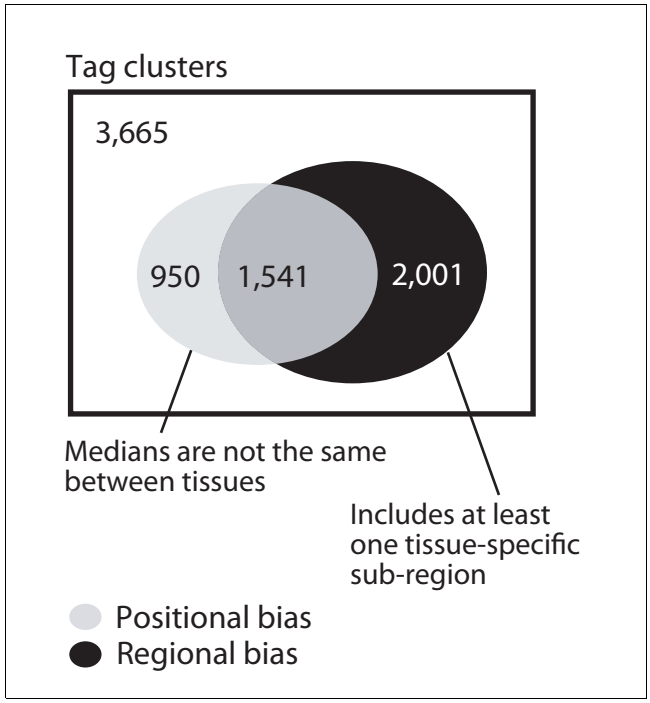

Figure 5
Tag clusters classified as positional and regional biases. The Venn diagram shows the number of tag clusters where start site selection is biased due to tissue. A biased tag cluster can either have distinct medians for tag starts for different tissues (termed positional bias) and/or have subregions that have a significantly different tag composition than the whole tag cluster. Most tag clusters that have positional bias also have subregions with significantly different tag composition. Only the tag clusters that are not classified as positional bias but include tissue-specific subregions are termed regional biases.

genomic sequencing revealed associations of DNA methylation with tissue-specific expression and partial methylation in some examples [36].

To explore the possibility that the fine-grained tissue specificities are associated with differential methylation, we compared these 150 differentially methylated regions identified by Song and coworkers [36] with our classification. Most of the regions are located at promoters and CpG islands, and 29 of the tag clusters used here overlap the differentially methylated regions. Of the 29 tag clusters, 13 (44%) and 11 (37%) are classified as positionally and regionally biased, respectively. These fractions are substantially larger than the fractions of all tag clusters (30% for positional bias and 25% for regional bias) and the fractions of CpG related tag clusters (34% for positional bias and 29% for regional bias), but additional data are required to prove the association with differential methylation rigorously. Given these initial results, we hypothesize that differences in DNA methylation due to cellular context is one of several mechanisms responsible for the observed difference in TSS selection between tissues.

Conclusion

We found that TSSs are tissue-specifically utilized within a tag cluster, rather than uniformly among tissues, in about half of all tag clusters in this study. Tag clusters with multiple and prominent CAGE tag peaks and positionally biased tissue specificity can be interpreted as distinct and overlapping promoters. On the other hand, a substantial number of tag clusters contain broad TSSs with regionally biased tissue specificity. Although detailed understanding of their regulation will require further experimentation, our comparisons with genome imprinting candidates raise the hypothesis that some of these tissue-specific TSS usages are regulated via DNA methylation and/or subsequent chromatin remodeling.

Our study is based on a limited number of 22 tissues profiled by CAGE, and the number of tag clusters with fine-grained tissue specificities is bound to increase when more tissues and conditions are added. Our results give rise also to questions

Table 1

Statistical significance of associations with CpG islands, CAGE tag shape classes, promoter expression, and imprinting candidates

		Positional bias	Regional bias	Other
CpG islands	CpG	4.11×10^{-25}*	1.65×10^{-56}*	1.00
	Not CpG	1.00	1.00	2.22×10^{-113}*
CAGE tag shape class	Single dominant peak	1.00	1.00	2.13×10^{-182}*
	Broad distribution with a dominant peak	1.92×10^{-02}	3.72×10^{-01}	9.89×10^{-01}
	Multimodal Distribution	2.37×10^{-12}*	1.49×10^{-02}	1.00
	General broad Distribution	6.70×10^{-01}	3.19×10^{-79}*	1.00
Imprinting candidates	Maternal imprinting candidates	7.94×10^{-01}	1.86×10^{-04}*	9.97×10^{-01}
	Paternal imprinting candidates	6.65×10^{-04}*	2.77×10^{-01}	1.00
	Not candidates	9.69×10^{-01}	9.99×10^{-01}	7.08×10^{-06}*

Shown are the levels of statistical significance (*P* values) of associations with CpG islands, CAGE tag shape classes, promoter expression, and imprinting candidates, where the Fisher's exact test is used for the null hypothesis that the two examined sets do not have any positive associations. *Statistically significant findings. CAGE, cap analysis of gene expression.

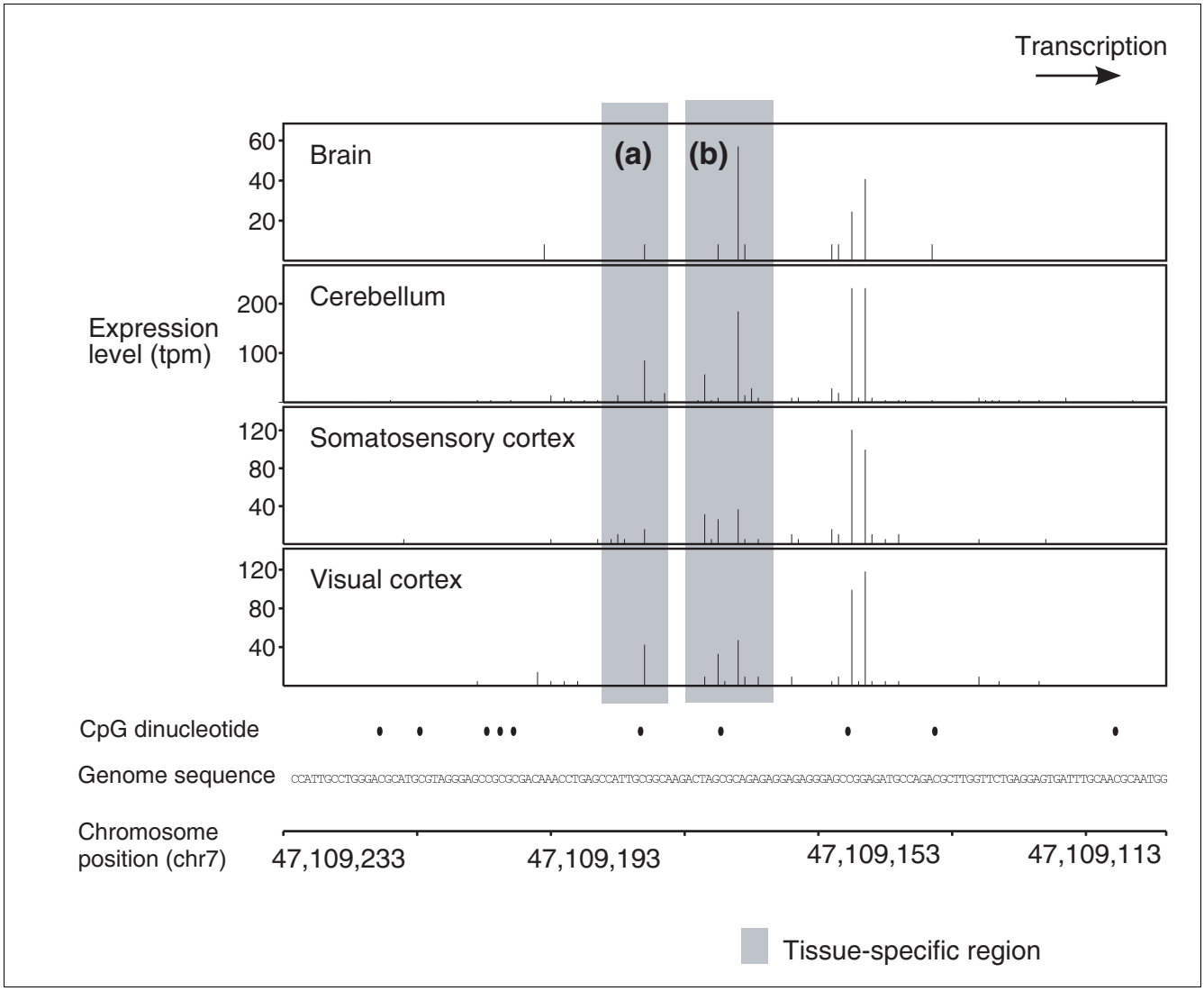

Figure 6
TSS usage and CpG dinucleotides at the 5' end of *Snrpn*. Expression levels of tag cluster T07R02CED41C, located at the 5' end of *Snrpn* on the reverse strand of chromosome 7, are shown. Only tissues with tpm (transcripts per million) above 100 and more than ten CAGE tags in the tag cluster are shown, and their expression levels are displayed as histograms. Gray backgrounds indicate regions with different tissue specificity from the remaining part of the tag cluster. Genome sequences are displayed below the graphs, and CpG sites are indicated by dots. CAGE, cap analysis of gene expression; TSS, transcription start site.

about TSSs and transcript 5' ends. In general, the transcripts with the most upstream 5' ends have been utilized to define TSSs of genes [37,38]. However, our findings imply that this methodology is biologically relevant only in some cases, because specific transcription starts frequently from nearby but distinct sites depending on tissue preferences. Comprehensive detection of TSSs in all tissues and conditions will be required to gain a complete understanding of transcriptional regulation and of the logic behind specific recruitment of transcription factors within core promoter elements.

This study highlights a property of core promoters that is little explored and less understood; it is clear that start site selection within promoters is a highly regulated process and that core promoters cannot be considered simply as standard templates serving to integrate signals from other *cis* regulatory elements.

Materials and methods

Data source

Mouse tag clusters and CAGE tag counts based on NCBI build 33 were retrieved from the CAGE Analysis Database [39], which provides CAGE tag counts for each library at the base pair level, associations of tag clusters with gene names, and additional information [40]. CAGE data belonging to different libraries from the same tissue were merged. Twenty-two tissues were used for our analysis (Table 2). Although some of

Table 2

Utilized tissues and their CAGE tag counts in mouse

Tissue	Total CAGE tags
Adipose	237,434
Amnion	381
Brain	122,603
Cerebellum	211,541
Cerebral cortex	17,510
Diencephalon	43,044
Embryo	1,077,740
Eye	1,246
Heart	34,743
Hippocampus	2,728
Liver	2,130,614
Lung	1,129,877
Macrophage	1,231,138
Mammary gland	1,294
Medulla oblongata	3,193
Muscle	42,294
Placenta	249
Prostate gland	57,057
Somatosensory cortex	190,762
Striatal primordia	39,693
Testis	109,150
Visual cortex	211,620

CAGE, cap analysis of gene expression.

them were not mutually exclusive, for example brain and cerebellum, they were treated as different categories. CpG island locations used in the above analysis are retrieved from the UCSC Genome Browser Database [41].

Regionally biased tissue specificity

Here, we aimed to test the null hypothesis that a tag cluster does not contain any internal regions with different tissue specificity from the remaining part. Although a large number of CAGE tags are used, some regions inside tag clusters have few tags, because of our tag cluster definition stating that any region with at least one tag is a part of a tag cluster. To achieve a reliable test in cases with such a small number of CAGE tags, we used the tissue specificity score (TS) and the negative log value of its relative change (rTS), which was devised for finding tissue-specificity of alternative splicing from EST libraries [23]. Bayesian statistics is used to make a reliable detection even among tissues with small numbers of ESTs.

Call an internal region in a tag cluster R_{int}, the remaining part R_{rem}, a tissue T, and all of the other tissues U. Let the hidden or true frequency of CAGE tag counts in R_{int} derived from T be $\theta_{int,T}$, and similarly for the other variables to yield $\theta_{int,U}$, $\theta_{rem,T}$ and $\theta_{rem,U}$. They are normalized and should fulfill the following equations:

$$\theta_{int,T} + \theta_{rem,T} = 1$$

$$\theta_{int,U} + \theta_{rem,U} = 1$$

The tissue specificity score (TS) is calculated from the observed CAGE tag counts as follows:

$$TS = 100\ (P\ [\theta_{int,T} > 0.5 \mid obs] - P\ [\theta_{rem,T} > 0.5 \mid obs])$$

Let the observed CAGE tag counts be $N_{int,T}$ in the internal region derived from T, and the negative log value of its relative change (rTS) is defined as follows:

$$rTS = -\log10\ (\Delta TS/TS)$$

Where $\Delta TS = |\ TS(N_{int,T}) - TS(N_{int,T} - 1)\ |$. A high TS score indicates that the internal region is much preferred in the tissue in comparison with the all of the other tissues, and a high rTS value indicates that the TS value is stable even if a single CAGE tag is not sequenced by chance.

This examination of the null hypothesis, that the internal region in the tag cluster does not exhibit different tissue specificity from the remaining part, is applied for each tissue. Each 21 bp subregion around a genome position where any CAGE tag alignment starts is tested, and the tested subregions can overlap. Because of this evaluation being conducted repeatedly in a tag cluster, we adopted more rigorous thresholds than were used in the original publication of this method, namely TS score above 90 and rTS score above 0.9.

Statistical test for associations

Associations of the fine-grained tissue specificities with CpG islands, shapes of CAGE tag distribution, and genome imprinting candidates were evaluated by one-sided Fisher's exact test. A 2 × 2 contingency table for two sets of tag clusters is constructed, and the *P* value for the null hypothesis that the two sets do not have any positive association is evaluated.

Additional data files

The following additional data are available with the online version of this paper. Additional file 1 is a document including all of our classifications of the tag clusters and their attributes in tab-delimited format.

Acknowledgements

We should like to thank G McLachlan for advice on statistics; B Lenhard, DA Hume, V Bajic, and SL Tan for useful discussions about promoter analysis; T Kasukawa for scientific and technical discussion; A Karlsson for English editing; K Nakano and H Murakami for building computational systems; K Yoshida and K Murata for support; and all members of the FANTOM consortium. This study was supported by Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to YH; Research Grant for Advanced and Innovational Research Program in Life Science to JK; a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan to YH; and a

Grant for the Strategic Programs for R&D of RIKEN to YH. MCF is a University of Queensland Postdoctoral Fellow.

References

1. Levine M, Davidson EH: **Gene regulatory networks for development.** *Proc Natl Acad Sci USA* 2005, **102:**4936-4942.
2. Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424:**147-151.
3. Suzuki Y, Taira H, Tsunoda T, Mizushima-Sugano J, Sese J, Hata H, Ota T, Isogai T, Tanaka T, Morishita S, *et al.*: **Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites.** *EMBO Rep* 2001, **2:**388-393.
4. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, *et al.*: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626-635.
5. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter.** *Annu Rev Biochem* 2003, **72:**449-479.
6. Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, *et al.*: **CAGE: cap analysis of gene expression.** *Nat Methods* 2006, **3:**211-222.
7. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, *et al.*: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.** *Proc Natl Acad Sci USA* 2003, **100:**15776-15781.
8. Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M, Hayashizaki Y: **Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes.** *Genome Res* 2000, **10:**1617-1630.
9. Carninci P, Waki K, Shiraki T, Konno H, Shibata K, Itoh M, Aizawa K, Arakawa T, Ishii Y, Sasaki D, *et al.*: **Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia.** *Genome Res* 2003, **13:**1273-1289.
10. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, *et al.*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309:**1559-1563.
11. Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, *et al.*: **Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation.** *Nat Methods* 2005, **2:**105-111.
12. Harbers M, Carninci P: **Tag-based approaches for transcriptome research and genome annotation.** *Nat Methods* 2005, **2:**495-502.
13. Coleman RA, Pugh BF: **Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA.** *J Biol Chem* 1995, **270:**13850-13859.
14. Hochheimer A, Tjian R: **Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression.** *Genes Dev* 2003, **17:**1309-1320.
15. Biel M, Wascholowski V, Giannis A: **Epigenetics: an epicenter of gene regulation: histones and histone-modifying enzymes.** *Angew Chem Int Ed Engl* 2005, **44:**3186-3216.
16. Futscher BW, Oshiro MM, Wozniak RJ, Holtan N, Hanigan CL, Duan H, Domann FE: **Role for DNA methylation in the control of cell type specific maspin expression.** *Nat Genet* 2002, **31:**175-179.
17. Jaenisch R, Bird A: **Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.** *Nat Genet* 2003:245-254.
18. Jones PA, Takai D: **The role of DNA methylation in mammalian epigenetics.** *Science* 2001, **293:**1068-1070.
19. Sutherland JE, Costa M: **Epigenetics and the environment.** *Ann N Y Acad Sci* 2003, **983:**151-160.
20. Wolffe AP, Matzke MA: **Epigenetics: regulation through repression.** *Science* 1999, **286:**481-486.
21. Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, Yamamoto J, Sekine M, Tsuritani K, Wakaguri H, *et al.*: **Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes.** *Genome Res* 2006, **16:**55-65.
22. Siegel S, Castellan NJ: *Nonparametric Statistics for the Behavioral Sciences* 2nd edition. New York: McGraw-Hill; 1988.
23. Xu Q, Modrek B, Lee C: **Genome-wide detection of tissue-specific alternative splicing in the human transcriptome.** *Nucleic Acids Res* 2002, **30:**3754-3766.
24. Wilkins JF: **Genomic imprinting and methylation: epigenetic canalization and conflict.** *Trends Genet* 2005, **21:**356-365.
25. Robertson KD: **DNA methylation and chromatin: unraveling the tangled web.** *Oncogene* 2002, **21:**5361-5379.
26. Reik W, Walter J: **Genomic imprinting: parental influence on the genome.** *Nat Rev Genet* 2001, **2:**21-32.
27. Ferguson-Smith AC, Reik W: **The need for Eed.** *Nat Genet* 2003, **33:**433-434.
28. Mager J, Montgomery ND, de Villena FP, Magnuson T: **Genome imprinting regulated by the mouse Polycomb group protein Eed.** *Nat Genet* 2003, **33:**502-507.
29. Nikaido I, Saito C, Wakamoto A, Tomaru Y, Arakawa T, Hayashizaki Y, Okazaki Y: **EICO (Expression-based Imprint Candidate Organizer): finding disease-related imprinted genes.** *Nucleic Acids Res* 2004, **32:**D548-D551.
30. Nikaido I, Saito C, Mizuno Y, Meguro M, Bono H, Kadomura M, Kono T, Morris GA, Lyons PA, Oshimura M, *et al.*: **Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling.** *Genome Res* 2003, **13:**1402-1409.
31. Mizuno Y, Sotomaru Y, Katsuzawa Y, Kono T, Meguro M, Oshimura M, Kawai J, Tomaru Y, Kiyosawa H, Nikaido I, *et al.*: **Asb4, Ata3, and Dcn are novel imprinted genes identified by high-throughput screening using RIKEN cDNA microarray.** *Biochem Biophys Res Commun* 2002, **290:**1499-1505.
32. Shemer R, Birger Y, Riggs AD, Razin A: **Structure of the imprinted mouse Snrpn gene and establishment of its parental-specific methylation pattern.** *Proc Natl Acad Sci USA* 1997, **94:**10267-10272.
33. Hajkova P, Erhardt S, Lane N, Haaf T, El-Maarri O, Reik W, Walter J, Surani MA: **Epigenetic reprogramming in mouse primordial germ cells.** *Mech Dev* 2002, **117:**15-23.
34. Hatada I, Hayashizaki Y, Hirotsune S, Komatsubara H, Mukai T: **A genomic scanning method for higher organisms using restriction sites as landmarks.** *Proc Natl Acad Sci USA* 1991, **88:**9523-9527.
35. Hattori N, Abe T, Hattori N, Suzuki M, Matsuyama T, Yoshida S, Li E, Shiota K: **Preference of DNA methyltransferases for CpG islands in mouse embryonic stem cells.** *Genome Res* 2004, **14:**1733-1740.
36. Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, Nagase H, Held WA: **Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression.** *Proc Natl Acad Sci USA* 2005, **102:**3336-3341.
37. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33:**D501-D504.
38. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: **The Ensembl automatic gene annotation system.** *Genome Res* 2004, **14:**942-950.
39. **CAGE Analysis Database** [http://fantom31p.gsc.riken.jp/cage_analysis/]
40. Kawaji H, Kasukawa T, Fukuda S, Katayama S, Kai C, Kawai J, Carninci P, Hayashizaki Y: **CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis.** *Nucleic Acids Res* 2006, **34:**D632-D636.
41. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, *et al.*: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31:**51-54.