Bioinformatic Analysis of Imprinted CpG Islands in *Mus Musculus*

Daniel Patrick Riordan
Churchill College
University of Cambridge
August 2003

This dissertation is submitted for the degree of Master of Philosophy.

PREFACE

This dissertation is my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

ACKNOWLEDGEMENTS

I wish to acknowledge my advisor, Richard Durbin, for his valuable guidance and direction in helping me to complete this work. The other members of the Durbin Lab – Ian Korf, Lachland Coin, and Ashwin Hajarnavis – also helped me in various ways throughout the year, for which I am thankful. I would also like to recognize Rachel Smith and Gavin Kelsey for providing the initial dataset upon which this analysis was based and for their helpful suggestions. I am grateful to have been supported by a Churchill Scholarship from the Winston Churchill Founation of the United States of America and also by the Wellcome Trust in carrying out this research. I would especially like to recognize all of my friends and family who have always been there for me. Most of all, I want to thank my parents, Kathleen and John Riordan, without whose constant love and support I wouldn't be the person I am today.

TABLE OF CONTENTS

ABSTRACT

Bioinformatic Analysis of Imprinted CpG Islands in *Mus musculus*

For some genes the maternally and paternally inherited alleles are differentially labeled with epigenetic markings leading to mono-allelic expression from only one copy, a process known as genomic imprinting. Previous studies have shown that imprinting is often associated with differential methylation of CpG islands (CGIs) and that SINE repeats occur less frequently in imprinted regions. This study identifies additional features of differentially-methylated CpG islands (DMR-CGIs) and uses them to help predict novel imprinted loci. A database containing sequences for 60 imprinted genes at 41 loci and 13,070 control genes in mouse was created and analyzed according to repeat content and CGI sequence properties. The SINE repeat content was significantly reduced at imprinted versus control loci, confirming previous reports. The sequence composition of CGIs associated with imprinted and control genes was also examined, and a considerable number of oligonucleotides with significantly different frequencies between DMR and control CGIs were found. A scoring function based on these oligonucleotides was developed that assigns greater scores to DMR-CGIs than controls at a highly significant level ($p < 10^{-5}$), and this scoring function was used in conjunction with regional SINE repeat content to predict novel imprinted loci. Genes associated with the predicted novel imprinted loci were compared with another set of candidate imprinted genes that were recently experimentally identified by large-scale expression profiling of FANTOM2 mouse cDNAs, and there was considerable overlap between both sets.

INTRODUCTION

Mammalian autosomal genes are normally present in duplicate - one copy is inherited from the mother and one copy is inherited from the father - and both copies of each gene are usually regulated in the same manner independently of the parent of origin. A notable exception to this general rule of genetic equality observed in classical Mendelian inheritance is the phenomenon of genomic imprinting. Imprinting is the process by which the maternally- and paternally-inherited copies of the same gene become epigenetically labeled in different ways, resulting in mono-allelic expression of that gene from only one locus. Imprinted genes are typically always transcribed from one allele and silenced at the other in a parent-specific fashion, although more complex expression patterns have also been observed, such as tissue-specific imprinting.

Although imprinting only occurs in a small minority of autosomal genes, estimated to be at least 100 in mouse (Reik and Walter, 2001), it is an important biological process that is essential for normal mammalian development. Disruption of imprinting has been implicated in a number of human genetic diseases, including Prader-Willi syndrome, Angelman syndrome, and Beckwith-Wiedemann syndrome, as well as different types of cancer (Reik and Walter, 2001). It has also been suggested that defects in imprinting could be involved in many neurological disorders that exhibit parent-of-origin effects, such as autism, bipolar affective disorder, epilepsy, and schizophrenia (Reik and Walter, 2001). Furthermore, epigenetic mechanisms underlying imprinted gene expression have also been shown to be involved in the regulation of non-imprinted genes during development, indicating that knowledge of imprinting could contribute to our appreciation of more general modes of mammalian gene regulation (Ehrlich, 2003).

Investigations into the mechanisms responsible for ‘writing’ genomic imprints have demonstrated that DNA methylation is a major epigenetic modification used to distinctly label the maternally- and paternally-inherited copies of imprinted genes. In mammalian genomes, DNA methylation occurs on

cytosines at 5'-CG-3' dinucleotides (CpG), and most CpG sites in the genome are normally methylated (Bird, 2002). Since methylated CpG mutates into TpG (or, equivalently, the reverse complement CpA) at a high rate, the overall observed frequency of genomic CpG is approximately 1%, which is roughly 25% of the rate expected by chance according to the single base frequencies and an assumption of independence (Bird, 2002). However, there are some GC-rich regions of the genome that exhibit elevated frequencies of CpG around 8%. These regions, known as CpG Islands (CGIs), were first identified experimentally as 'HpaII-small fragments' because of their increased sensitivity to digestion by HpaII, a restriction enzyme which cleaves specifically at non-methylated restriction sites containing CpG (Bird, 1986). However, CGIs are now typically identified computationally based on their sequence properties, conventionally defined as regions of 200bp or more with GC-content ≥ 50% and an observed/expected CpG ratio ≥ .60 (Gardiner-Garden and Frommer, 1987). CGIs are often associated with the 5' ends of genes, and most CGIs are usually unmethylated, which accounts for their elevated rate of CpG (Bird, 1986). In contrast, many CGIs associated with imprinted genes are the sites of differential DNA methylation on only one chromosome that marks imprinted loci differently according to the parent of origin (Reik and Walter, 2001).

The importance of CGIs for imprinting has been demonstrated through a number of studies in mice and humans. Deletion of CGIs has been shown to disrupt imprinted gene regulation at several loci, including *H19/Igf2*, the PWS/AS locus, the BWS locus, and *Igf2r/Air*, indicating a central role for CGIs in imprinting (Bestor, 2000). Experiments involving bisulfite sequencing and methylation-sensitive restriction enzymes have clearly shown that differential methylation of CGIs occurs at imprinted loci (Reik and Walter, 2001). The functional importance of methylation has been established by analysis of mouse knockout mutants lacking *Dnmt1*, the main maintenance methyltransferase, and *Dnmt3L*, which encodes a non-catalytic protein in the DNA methyltransferase family with sequence similarity to the *de novo* methyltransferases *Dnmt3a* and *Dnmt3b*. Mouse embryos defective in *Dnmt1* display a loss of proper imprinted

gene expression and die in mid-gestation, implying that DNA methylation is necessary to maintain the epigenetic status of maternal and paternal genes at imprinted loci (Li *et al*, 1993). Embryos from *Dnmt3L*-/- female mice lack differential methylation of maternal DNA at imprinted regions (whereas global DNA methylation remains unperturbed) and die soon after implantation with evident growth abnormalities (Bourc'his *et al*, 2001). Together, these investigations demonstrate that genomic imprints can be established by germline-specific methylation of CGIs in imprinted regions which differentially marks the paternal and maternal copies of imprinted genes.

Imprinted genes have typically been identified in a variety of ways (Reik and Walter, 2001). Some individual imprinted genes have been fortuitously discovered by knockout experiments of genes that subsequently displayed parent-specific expression. Imprinted genes have also been identified based on their location near other known imprinted genes or in chromosomal regions associated with imprinting phenotypes.

However, there are two main types of screens that have been used to systematically search for imprinted genes, both of which usually rely on comparisons between uniparental embryos containing either only paternally-derived chromosomes (androgenetic embryos) or only maternally-derived chromosomes (parthenogenetic embyos). The first type of screen searches for differences in DNA methylation patterns between androgenetic and parthenogenetic embryos after digestion with methylation-sensitive restriction enzymes using techniques such as representational difference analysis or restriction landmark genome scanning. The second type of screen for identifying imprinted genes is based on comparisons of cDNA levels between uniparental embryos of maternal and paternal origin. Recently, a large-scale screen of this type was used to identify 2,114 candidate imprinted genes by using microarrays to detect differential expression of 27,663 FANTOM2 mouse cDNAs between the total tissue of 9.5 day old parthenogenetic and androgenetic embryos (Nikaido *et al*, 2003). Although bioinformatic techniques have not been previously used to discover novel imprinted genes, previous research efforts have aimed to identify

sequence characteristics unique to imprinted regions in the human genome.

Significant differences in the guanine and cytosine, CpG, and repeat content of imprinted regions relative to control loci were recently observed in two separate studies (Greally, 2002 and Ke *et al*, 2002), both of which were based on examination of the features of large sequence windows, typically 50kb in size, spanning imprinted loci. The most striking observation was a marked decrease in the content of short interspersed transposable elements (SINEs) near imprinted genes. A possible model explaining this is that SINE accumulation in imprinted regions was selected against because non-specific methylation of SINEs could deleteriously interfere with parent-specific methylation of imprinted loci (Greally, 2002). Although direct repeat sequences have also been historically described to correlate with imprinted regions, the generality of these observations remains in question (Okamura *et al*, 2000 and Arnaud *et al*, 2003). However, no previous published investigations have directly examined the sequences of imprinted CGIs themselves in order to explore the possibility that CGIs at imprinted loci could differ in composition from control CGIs at non-imprinted loci.

Based on the clear importance of CGIs to genomic imprinting, it seems reasonable to hypothesize that significant differences in the sequence composition of imprinted and control CGIs could be detected. There are two main reasons that imprinted CGIs might differ in sequence composition from control CGIs. First, the fact that only one copy of imprinted CGIs is unmethylated might lead to observable differences in nucleotide composition when compared to CGIs for which both copies are unmethylated; after all, the normally unmethylated status of most CGIs is believed to explain their unique sequence features of GC- and CpG-richness. Secondly, it is also possible that the distribution of binding sites for trans-acting factors involved in epigenetic regulation could be enriched or reduced within imprinted CGIs, leading to detectable differences in their sequence composition. Indeed, *CTCF*, an 11-zinc finger transcription factor which regulates chromatin boundaries and may act as both a repressor and activator, has been shown to be involved directly in imprinted regulation of the *Igf2/H19* locus by binding to a differentially methylated CGI region (Schoenherr *et al*, 2003).

Additional CTCF-binding motifs have also been identified within differentially-methylated domains at other loci (Kim *et* al, 2003 and Hikichi *et* al, 2003) and it is possible that other trans-acting factors, such as those involved in the establishment of imprints, may bind specifically to imprinted CGIs to mediate epigenetic regulation of imprinted genes. We therefore decided to analyze the sequence features of CGIs from imprinted and control loci in mouse in order to look for significant characteristics which may be involved in genomic imprinting.

MATERIALS & METHODS

CpG Island Identification

CGIs are GC-rich DNA sequences that exhibit an elevated frequency of CpG dinucleotides and are often unmethylated and associated with the 5' ends of genes. To identify CGIs in mouse, the gene sequence and upstream region (50kb from the annotated start of the gene) for all 13,112 autosomal mouse genes that were defined by EnsEMBL as 'known genes' were extracted from the mus_musculus_core_9_3 database of December 2002 (www.ensembl.org). All X-linked genes were excluded from the analysis because the process of X chromosome inactivation shares many features with genomic imprinting, including epigenetic regulation and differential methylation of CpG islands (Lee, 2003). CGIs were identified in autosomal sequences using the cpgplot program (http://www.emboss.org) with default parameters. The imprinted genes *H19*, *Peg3*, and *Snrpn* all contain known differentially methylated CGIs which could not be detected using default parameters, so an alternative minimum length parameter (minlen=100) was used for these genes. Nearby cpgplot-identified islands were merged together to form a single CGI if the overall sequence properties of the resulting island satisfied the conventionally accepted criteria for CGIs (Length ≥ 200bp, GC-content ≥ 0.5, CpG Obs/Exp ratio ≥ 0.6). Duplicate and overlapping islands were then removed to yield a set of 13,665 unique CGIs.

Additional CTCF-binding motifs have also been identified within differentially-methylated domains at other loci (Kim *et* al, 2003 and Hikichi *et* al, 2003) and it is possible that other trans-acting factors, such as those involved in the establishment of imprints, may bind specifically to imprinted CGIs to mediate epigenetic regulation of imprinted genes. We therefore decided to analyze the sequence features of CGIs from imprinted and control loci in mouse in order to look for significant characteristics which may be involved in genomic imprinting.

MATERIALS & METHODS

CpG Island Identification

CGIs are GC-rich DNA sequences that exhibit an elevated frequency of CpG dinucleotides and are often unmethylated and associated with the 5' ends of genes. To identify CGIs in mouse, the gene sequence and upstream region (50kb from the annotated start of the gene) for all 13,112 autosomal mouse genes that were defined by EnsEMBL as 'known genes' were extracted from the mus_musculus_core_9_3 database of December 2002 (www.ensembl.org). All X-linked genes were excluded from the analysis because the process of X chromosome inactivation shares many features with genomic imprinting, including epigenetic regulation and differential methylation of CpG islands (Lee, 2003). CGIs were identified in autosomal sequences using the cpgplot program (http://www.emboss.org) with default parameters. The imprinted genes *H19*, *Peg3*, and *Snrpn* all contain known differentially methylated CGIs which could not be detected using default parameters, so an alternative minimum length parameter (minlen=100) was used for these genes. Nearby cpgplot-identified islands were merged together to form a single CGI if the overall sequence properties of the resulting island satisfied the conventionally accepted criteria for CGIs (Length ≥ 200bp, GC-content ≥ 0.5, CpG Obs/Exp ratio ≥ 0.6). Duplicate and overlapping islands were then removed to yield a set of 13,665 unique CGIs.

Smith & Kelsey provided a curated database containing 60 mouse genes that are known to be imprinted. The known imprinted genes represented in the EnsEMBL database were derived from a total of 41 unique gene loci, since some genes (e.g. *Copg2* and *Copg2as*) were associated with the same EnsEMBL gene identifier. 45 of the unique CGIs identified were associated with the 41 imprinted gene loci, and these CGIs were categorized according to their methylation status. The 27 CGIs which coincided with known differentially-methylated regions were classified as DMR-CGIs, and the additional 18 CGIs (which were unmethylated, methylated on both alleles, or of unkown methylation status) were classified as UMR-CGIs. The remaining 13,619 autosomal CGIs that were not associated with known imprinted genes were classified as 'control CGIs.'

Significant *K*-mer Analysis

To identify sequences that were significantly enriched or depleted in imprinted CGIs relative to control CGIs, the frequencies of all 4^k possible *k*-mers in the set of imprinted CGIs were calculated and compared to the frequencies in the control set for a range of values of *k*. Both strands of each sequence were considered, and *k*-mers containing the ambiguity code N were excluded from the analysis. The statistical significance of observed differences between frequencies in the imprinted and control sets was then determined for each *k*-mer by calculating an exact *p*-value according to a Poisson distribution with rate parameter λ defined as the frequency of the *k*-mer in control CGIs, which is equal to the maximum-likelihood estimate of λ. For a given *k*-mer that occurs *n* times in an imprinted set with *T* total *k*-mers, the *p*-value was calculated as the probability of having observed *n* or fewer occurrences in the imprinted CGIs according to the null hypothesis that the distributions for that *k*-mer are identical in the imprinted and control sets. The *p*-value is then given by the expression $\sum_{x=0}^{n} e^{-(\lambda T)} \times (\lambda T)^x / x!$. *K*-mers with *p*-values less than or equal to α or greater than or equal to (1-α) were defined as 'Significant *k*-mers' for a range of significance levels ($\alpha = 10^{-2}, 10^{-3},$

10^{-4}, 10^{-5}).

CGI Scoring Function

A log-odds scoring function *S* based on the significant *k*-mers was defined as $S(X) = 10 \times \sum_{j=1}^{L-k+1} I_{sig}(x_j) \times \log_2\left(\frac{f_{impr}(x_j)}{f_{ctrl}(x_j)}\right)$ where *X* is a CGI sequence to be scored with length *L*, x_j is the *k*-mer starting at position *j* in *X*, $f_{impr}(x_j)$ and $f_{ctrl}(x_j)$ are the frequencies of x_j in the imprinted and control CGI sets, respectively, and $I_{sig}(x_j)$ is an indicator function equal to one if x_j is a 'significant *k*-mer' and equal to zero otherwise. Thus, the scoring function S depends on three sets of parameters – the *k*-mer frequencies in an imprinted dataset, the *k*-mer frequencies in a control dataset, and a set of significant *k*-mers whose frequencies are significantly different in the imprinted and control datasets. The complete sets of DMR and control CGIs were used to calculate these parameters for scoring UMR and control CGIs. However, for scoring the DMR-CGIs it was necessary to take extra measures in order to avoid over-fitting due to the small sample size of imprinted sequence data. Therefore, a 'jack-knife' approach was adopted, whereby the parameters used to score each imprinted CGI were calculated based on an adjusted dataset containing all imprinted CGIs **except** the one being scored, as well as the entire set of control CGIs. Without this adjustment to the scoring procedure, spuriously "significant" results can be misleadingly obtained.

RESULTS

Dataset Properties

A database of 60 known imprinted mouse genes provided by Smith and

10^{-4}, 10^{-5}).

CGI Scoring Function

A log-odds scoring function *S* based on the significant *k*-mers was defined as $S(X) = 10 \times \sum_{j=1}^{L-k+1} I_{sig}(x_j) \times \log_2\left(\frac{f_{impr}(x_j)}{f_{ctrl}(x_j)}\right)$ where *X* is a CGI sequence to be scored with length *L*, x_j is the *k*-mer starting at position *j* in *X*, $f_{impr}(x_j)$ and $f_{ctrl}(x_j)$ are the frequencies of x_j in the imprinted and control CGI sets, respectively, and $I_{sig}(x_j)$ is an indicator function equal to one if x_j is a 'significant *k*-mer' and equal to zero otherwise. Thus, the scoring function S depends on three sets of parameters – the *k*-mer frequencies in an imprinted dataset, the *k*-mer frequencies in a control dataset, and a set of significant *k*-mers whose frequencies are significantly different in the imprinted and control datasets. The complete sets of DMR and control CGIs were used to calculate these parameters for scoring UMR and control CGIs. However, for scoring the DMR-CGIs it was necessary to take extra measures in order to avoid over-fitting due to the small sample size of imprinted sequence data. Therefore, a 'jack-knife' approach was adopted, whereby the parameters used to score each imprinted CGI were calculated based on an adjusted dataset containing all imprinted CGIs **except** the one being scored, as well as the entire set of control CGIs. Without this adjustment to the scoring procedure, spuriously "significant" results can be misleadingly obtained.

RESULTS

Dataset Properties

A database of 60 known imprinted mouse genes provided by Smith and

Kelsey was compared with the EnsEMBL mouse database and these known imprinted genes were mapped to 41 EnsEMBL gene loci (see Methods). Mouse CpG islands (CGIs) were identified in the gene sequences and upstream regions of these 41 imprinted mouse gene loci and the remaining 13,071 distinct autosomal gene loci annotated in EnsEMBL. CGIs were defined as sequences of at least 200bp with GC-content ≥ 50% and an observed/expected ratio of CpG ≥ 0.60. From imprinted loci, 45 unique CGIs were identified and 33 imprinted gene loci (80.5%) contained at least one CGI in their gene sequence or upstream region (Table 1). The imprinted CGIs were classified according to their methylation status: the 27 imprinted CGIs which coincide with known differentially-methylated regions were categorized as DMR-CGIs, and the remaining 18 CGIs were termed UMR-CGIs (Table 1). The set of DMR-CGIs was considered to represent 'imprinted CGIs' for the purpose of this study.

An additional 13,619 unique 'control CGIs' were identified, and 10,393 of the control gene loci (79.5%) contained at least one CGI in their gene sequence or upstream region. The number of control genes associated with CGIs in our dataset is higher than the rate of ~50% typically reported (Reik and Walter, 2001), but this discrepancy can be attributed to a difference in the minimum length criterion used for CGI definition (200bp vs. 500bp). When a minimum CGI length of 500bp is required, the number of control genes with CGIs is reduced to 7,393 (56.6%), consistent with previous reports. However, some of the imprinted CGIs which are known to be differentially methylated are also shorter than 500bp, so the less stringent minimum length requirement of 200bp was retained.

Repeat Element Content

Because mammalian CpG islands have been shown to vary in structure, the repetitive element content of the identified CGIs was initially examined. A small subset of 597 of the control CGIs (4.4%) and one of the imprinted CGIs (2.2%) were found to contain one or more RepeatMasker-identified SINE sequences annotated in EnsEMBL. This is consistent with the observation that

CGIs located near the transcription start site of genes are unlikely to be due to repeated sequences (Ponger *et al*, 2001) and indicates that the CGI sequences are not dominated by the presence of repetitive elements.

The distributions of different classes of repetitive elements in larger sequence windows extending beyond the CGIs in imprinted regions were then examined because they had been previously reported to differ significantly from control loci in humans (Greally, 2002 and Ke *et al*, 2002). To assess whether imprinted mouse loci also exhibit this property, the region containing each imprinted CGI and its 100kb flanking sequence (50kb upstream and 50kb downstream) was analyzed for its repeat content, defined as the percentage of bases in the region that are annotated as RepeatMasker-identified repeat sequences in EnsEMBL. The ten different repeat classes that were considered are Type I Transposons/LINE, Type I Transposons/SINE, Type II Transposons, Low Complexity regions, LTRs, RNA repeats, Satellite repeats, Simple repeats, Other/Y-chromosomal repeats, and Other repeats.

The repeat content distribution for each class was compared between imprinted and control regions by a Wilcoxon rank-sum test, and the average SINE content in imprinted regions of 7.4% was found to be significantly lower than the average SINE content of 14.4% in control regions ($p < 10^{-9}$). This reduction in SINE content has been previously noted for imprinted human genes and is hypothesized to reflect an active selection against SINE accumulation, presumably because SINEs may attract and spread non-specific methylation which could disrupt genomic imprinting (Greally, 2002). Our results demonstrate that this characteristic feature of human imprinted domains is also conserved in mouse.

Interestingly, paternally-methylated DMR-CGIs appear to have slightly higher SINE content on average (8.4%) than maternally-methylated DMR-CGIs (5.8%), and this result appears to be significant ($p < 0.05$). While this may reflect a difference in the selective forces acting at maternally and paternally methylated loci, it remains to be seen whether this initial finding will be supported by the discovery of additional DMR-CGIs and examination of their SINE content.

Although significant differences in the distribution of additional types of repeat sequences (e.g. Low-complexity repeats) between imprinted and control regions were also reported in previous studies, no other significant differences were observed for our dataset. This may signal that those features of imprinted human loci are not conserved in mouse, but this discrepancy could also be accounted for by differences in the sequence windows, analysis software, and repeat element classifications used in the analyses.

CpG Content of Imprinted CGIs

We next chose to focus on the sequence properties of the imprinted CGIs themselves. An intriguing hypothesis raised in previous investigations was the idea that differential methylation of CGIs at imprinted loci throughout half of their evolutionary history would be reflected by an erosion of their CpG content (Greally, 2002). This theory was not supported by those analyses, however, which failed to detect a significant reduction in the rate or number of CGIs occurring in sequence windows of varying length spanning imprinted domains (Greally, 2002 and Ke *et al*, 2002). As our dataset of CGIs presented an opportunity to clearly test this hypothesis, we initially compared the CpG content of DMR and control CGIs for any significant differences.

The number of CpG dinucleotides present in DMR-CGIs (6.76%) was less than the number of occurrences in the set of control CGIs (8.72%) at a highly significant level ($p < 10^{-5}$). This result strongly supports the hypothesis that differential methylation of CGIs at imprinted loci leads to a reduction in their CpG content.

Reinforcing this view, the number of TpG dinucleotides was shown to be significantly increased ($p < 10^{-6}$) in DMR-CGIs (6.38%) with respect to control CGIs (5.91%), consistent with the idea that mutation of methyl-CpG to TpG is responsible for the decrease in CpG content of DMR-CGIs. In contrast, the UMR-CGIs associated with imprinted genes did not significantly differ in TpG content (5.64%) and displayed a slight increase of CpG content (9.67%) in

comparison to control CGIs which was statistically significant ($p < 10^{-5}$), hinting that the UMR-CGIs may be compositionally distinct from DMR-CGIs. It is interesting to note that the reduction in CpG content of DMR-CGIs is not accompanied by a significant decline in the occurrence of CGIs at imprinted loci, suggesting that selection for the maintenance of CGIs which serve as functionally important sites of differential methylation for genomic imprinting is balanced against the mutational decay of CpG sites in these DMR-CGIs.

Significant *K*-mer Analysis

The finding that the rate of CpG and TpG dinucleotides varies significantly between DMR and control CGIs also raised the possibility that other significant differences in composition could be identified between DMR and control CGIs which may be functionally relevant to the process of genomic imprinting. Likewise, the fact that UMR-CGIs do not share these properties with DMR-CGIs suggested that UMR-CGIs may be surpisingly similar in composition to control CGIs, despite their location in imprinted domains and proximity to nearby DMR-CGIs. To further explore these issues, we next sought to explicitly identify other *k*-mers (oligonucleotide 'words' of length k) that were significantly enriched or reduced in DMR-CGIs and UMR-CGIs relative to control CGIs.

The distribution of all *k*-mers in DMR-CGIs and UMR-CGIs were compared to control CGIs in order to identify significant *k*-mers for a range of word lengths (k = 5, 6, 7) and significance levels ($\alpha = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$). For all combinations, the number of *k*-mers with significant differences between the DMR-CGIs and controls was dramatic, indicating that these two sets vary considerably in sequence composition (Table 2). This observed compositional heterogeneity is all the more striking considering that both sets of sequences were equally constrained to satisfy the sequence criteria of CpG islands.

On the other hand, between the UMR-CGIs and controls there were far fewer differences and the number of significant *k*-mers was only somewhat greater than would be expected merely by chance due to the large number of *k*-

mers tested. As there are fewer UMR-CGIs (18) than DMR-CGIs (27), the resulting diminution of the statistical power could be expected to lead to a slight decrease in the number of significant k-mers found. However, it is unlikely that this minor difference in sample size could account for the substantial reduction in the number of significant k-mers observed between UMR-CGIs and DMR-CGIs versus controls.

This result demonstrates that UMR-CGIs, unlike DMR-CGIs, are not strikingly different in composition from control CGIs, although both DMR-CGIs and UMR-CGIs are associated with the same imprinted domains. The UMR-CGIs can therefore effectively function as regional controls for the DMR-CGIs, indicating that the majority of significant differences observed between DMR and control CGIs are likely to be specifically related to differential methylation and not due to secondary effects, such as regional genomic characteristics of imprinted loci.

Clustering Analysis

In an effort to identify discernible sequence motifs which may be involved in differential methylation of imprinted loci, we clustered the set of heptamers (oligonucleotides of length 7) that had frequencies which significantly differed ($p < 10^{-5}$) in DMR and control CGIs. The 82 heptamers with different frequencies in the DMR and control CGI sets at the significance level $\alpha = 10^{-5}$ were clustered based on their sequence similarity. A pairwise dissimilarity distance was defined as the number of shifts + mismatches in the best global alignment of the two heptamers, and this distance was used for standard hierarchical average linkage clustering of the heptamers. A distance cutoff of 2.2 was chosen as this value represented the midpoint of the range of dissimilarity values over which the clusters remained stable for the longest duration (excluding the endpoints). Using this cutoff level, six clusters were defined that each consisted of four or more heptamers and collectively included almost half (38 / 82) of all the heptamers. As two of these clusters were the reverse strand equivalents of others, four unique

clusters were identified overall. Cluster-1 contained heptamers with frequencies that were both significantly increased and significantly decreased in DMR versus control CGIs, so these heptamers were divided to create Cluster-1A (decreased) and Cluster-1B (increased), yielding a total of five clusters. These five clusters were then multiply aligned using CLUSTALW (ref) and motifs corresponding to the alignments were generated with the Pictogram program (http://genes.mit.edu/pictogram.html).

This clustering analysis organized a subset of heptamers that occur at significantly different rates between DMR and control CGIs into aligned groups of similar sequences which are represented by motifs in order to facilitate the recognition of biologically meaningful patterns (Figure 1). For Clusters 2-4, no obvious similarity to known sequence motifs that are relevant to imprinting was readily apparent. However, examination of Motif-1A and Motif-1B revealed a noticeable similarity to CpG-rich CTCF-binding sites, which was confirmed by further inspection.

CTCF is a highly conserved protein with roles in gene activation, repression, silencing, and chromatin insulation which has been shown to bind to a wide range of extremely divergent ~50 bp target sites through differential use of its 11-zinc finger domains. Since 15 CTCF target site sequences (footprints defined by protection against DNaseI attack) with annotated CTCF-contacting guanines (determined by dG-methylation interference within CTCF-bound regions) were available (Ohlsson *et al*, 2001), we compared Motif-1A and Motif-1B to the heptamers overlapping CTCF-contacting guanines within these target sites in order to assess whether the motifs were likely to represent CTCF-binding sites. Three heptamers from Cluster-1A (CGCCGCC, CGCCGCG, CGCCGCG) mapped to CTCF-contacting guanine sites within CTCF target sites for chicken *MYC-FpV*, human *PIM-1* oncogene, and human *APP*, and one heptamer (TGCCGCG) from Cluster-1B also coincided with a CTCF-contacting guanine site within the footprint for DMD7 of the mouse *Igf2/H19* imprinting control region (an imprinted DMR-CGI represented in our dataset).

From this comparison it was apparent that Motifs-1A/1B may represent

binding sites for CTCF. However, it was somewhat surprising that CTCF-binding sites in Cluster-1A occurred less frequently in DMR-CGIs than controls, since *CTCF* is known to maintain differential methylation and regulate imprinted gene expression through binding at the *Igf2/H19* locus and is thought to act similarly at other imprinted loci (Schoenherr *et al*, 2003). This prompted us to examine whether the Cluster-1A heptamers were enriched within DMR-CGIs compared to the distribution that would be expected based on the marginal dinucleotide frequencies in DMR-CGIs and an assumption of independence at each position. When this was evaluated using a X^2 goodness-of-fit test, two heptamers (CCGCCGC, GCCGCCG) were found to differ significantly ($p < 10^{-2}$) from the distribution expected by dinucleotide marginals, and both displayed significant enrichment of *CTCF*-binding sites within the DMR-CGIs. This enrichment of *CTCF*-binding sites within DMR-CGIs agrees with the demonstrated importance of *CTCF* in regulation of imprinting via methylation-sensitive recognition of binding sites. Indeed, the consideration that words from Cluster-1B may be obtained by substituting TpG/CpA for CpG sites in words from Cluster-1A (e.g. **CG**CCGCG > **TG**CCGCG) may reflect a greater degree of methylation at these sites, where partial methylation could allow *CTCF*-binding and lead to disruption of imprinting.

However, the fact that Cluster-1A sites are enriched within DMR-CGIs but nevertheless are less frequent in DMR-CGIs than controls suggests that *CTCF*-binding sites are vastly enriched within control CGIs to an even greater extent. This was confirmed, as all of the Cluster-1A heptamers are found to occur in control CGIs at highly significant ($p < 10^{-8}$) levels greater than expected by the dinucleotide distributions within control CGIs. This indicates that, in addition to its well-documented involvement in maintaining differential methylation and regulating imprinted gene expression, *CTCF* could also play a central but currently under-appreciated role in the maintenance or establishment of CGIs in general. This possibility was previously alluded to based on the CpG-richness of *CTCF* target sites (Ohlsson *et al*, 2001) and would be consistent with the ubiquitous expression of *CTCF* as well as its ability to protect the maternal

Igf2/H19 locus from methylation via DNA-binding (Schoenherr *et al*, 2003).

These results demonstrate that the involvement of *CTCF* in genomic imprinting is reflected in the significantly different rates at which its binding sites occur between DMR and control CGIs and strongly suggest a more general role for *CTCF* associated with all CGIs. At the same time, these analyses account for a small yet biologically interesting subset of the many compositional differences that were observed between DMR and control CGIs. A full list of the significant heptamers (at the level $\alpha = 10^{-5}$) that were used in the clustering analysis with the frequency in DMR-CGIs and Control CGIs, the log odds ratio of the frequencies and the associated *p*-value for each heptamer is included in Appendix 1.

Imprinted CGI Prediction by Significant *K*-mer Composition

Although we were not able to explain the remaining significant sequence differences between DMR and control CGIs by known biological binding sites, the extent of these differences raised the possibility that this compositional variability could provide information for discrimination between imprinted and control CGIs and be used to facilitate the discovery of novel imprinted CGIs. To explore this possibility, we developed a log-odds scoring function which assigns higher scores to CpG islands that are more similar to DMR-CGIs than control CGIs in their composition of significant *k*-mers (see Methods).

All CGIs were scored using every combination of *k*-mer lengths ($k = 5, 6, 7$) and significance levels ($\alpha = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$) and the results are described in Table 3. In order to avoid over-fitting due to the small sample size of the DMR dataset, we scored each DMR-CGI in turn based on the parameters obtained using all DMR-CGIs other than itself, whereas the control and UMR-CGIs were scored based on the complete datasets. The CGI scores for DMR and control CGIs were compared and the DMR-CGIs were found to score higher than control CGIs at extremely significant levels.

Although the clustering analysis of individual heptamers described in the previous section was performed on k-mers significant at the level $\alpha = 10^{-5}$, the

greatest difference in mean CGI scores for DMR and control CGIs was obtained with $k = 7$ and $\alpha = 10^{-2}$. This difference was highly significant ($p = 3.7e\text{-}6$) and more than half of the DMR-CGIs (52%) received scores greater than the vast majority (95%) of all control CGIs. Although some 7-mers significant at the level $\alpha = 10^{-2}$ are likely to be 'false positives,' the fact that this level was optimal for scoring indicates that 7-mers with marginally significant differences can still be informative and contribute to prediction of imprinted CGIs. For this combination of *k*-mer length and significance level ($k = 7$, $\alpha = .01$) the cumulative score distributions were plotted and compared for DMR, UMR, and control CGIs (Figure 2) and these scores were used for all subsequent analyses.

The fact that CGI scores for DMR-CGIs were markedly greater than for controls clearly demonstrates that nucleotide compositional differences can effectively contribute to the identification of imprinted loci. No differences in CGI scores between the maternally and paternally methylated subsets of DMR-CGIs were apparent. However, significant differences in the CGI score distribution of UMR-CGIs and Control CGIs were also observed. The average score for the UMR-CGIs of –62 was significantly lower than for DMR-CGIs ($p < 10^{-4}$) and was even significantly lower than for control CGIs ($p = .013$), supporting the idea that UMR-CGIs are quite distinct in composition from DMR-CGIs. This result also indicates that there are informative features of imprinted sequences represented in the CGI scoring function which are independent of regional repeat content, since the UMR-CGIs are located in the same SINE-poor regions as DMR-CGIs but still receive very low scores.

To assess whether CGIs that appear likely to be imprinted based on the scoring function also exhibit other properties characteristic of imprinted loci, we examined the SINE content of the control CGIs with the 50 highest CGI scores (HS-CGIs). As imprinted loci characteristically display lower SINE content than control loci, a reduction in the SINE content of regions surrounding the HS-CGIs would provide further evidence to support the possibility that they could represent true imprinted loci. The distribution of SINEs in the regions containing the HS-CGIs and 100kb of flanking sequence was analyzed, and the average SINE

content of 12.6% for the HS-CGIs was significantly lower ($p < 0.05$) than the rate of 14.4% for other control CGIs (Figure 3). This demonstrates that CpG islands that are compositionally similar to DMR-CGIs also display a reduction in local SINE content which is a hallmark of imprinted loci, and is consistent with the idea that the HS-CGI set may contain novel imprinted CGIs. However, the SINE content of HS-CGIs is also significantly higher than the rate of 7.4% for the DMR-CGIs ($p < 0.001$). Therefore, the set of HS-CGIs is likely to contain a mixture of novel imprinted CGIs together with non-imprinted loci. This result suggested that information about regional SINE repeat content could be used in combination with nucleotide composition to improve prediction of imprinted loci.

Prediction of Imprinted CGIs using CGI Score & SINE Content

To predict novel imprinted loci using the CGI scores in conjunction with regional SINE repeat content, we developed a method that classifies CGIs as imprinted or non-imprinted by linear discriminant analysis. Incorporating the information represented by these two significant features of imprinted loci, a linear discriminant function that minimizes the mean squared error of classification was determined using the R statistical software package (www.R-project.org), and the results of classification are depicted in Figure 4. A threshold was chosen corresponding to a prediction region that correctly classifies 9 DMR-CGIs (33.3%) along with 249 control CGIs as imprinted loci. An additional requirement was imposed that all CGIs classified as imprinted must have greater CGI scores and lower SINE content than the respective median values for all control CGIs. This final classification yielded 9 DMR-CGIs and 218 control CGIs that are predicted to be candidate novel imprinted loci, which represents a 20-fold enrichment ($9/227 > 20 \times 27/13646$) of DMR-CGIs over the original dataset. Analogous classification schemes of equal sensitivity based only on CGI scores alone or SINE content alone achieve 10.2-fold and 5.8-fold enrichment, respectively. This demonstrates that CGI scores and SINE content are each powerful predictors of imprinting status, and that consideration of both of these

features in combination enhances prediction of imprinted loci.

Comparison to FANTOM2 Candidate Imprinted Transcripts

To assess whether genes associated with these predicted novel imprinted loci exhibit expression patterns indicative of genomic imprinting, we compared our dataset with another list of candidate imprinted genes that were recently identified by large-scale expression profiling of FANTOM2 mouse cDNA clones (Nikaido *et al*, 2003). The FANTOM2 set contained 1,958 autosomal mouse transcripts (and X-linked transcripts that were excluded from this analysis) which were predicted to be imprinted based on comparison of mRNA levels in uniparental mouse embryos of maternal and paternal origin. Control EnsEMBL genes corresponding to the FANTOM2 candidates were identified by sequence similarity using MEGABLAST 2.2.6 [with option –p 0.99 and default parameters otherwise] (ref) and requiring consistent chromosomal locations from both sets. 1,031 (53%) of the FANTOM2 transcripts were mapped in this way to 945 different EnsEMBL genes in our dataset with 1,697 control CGIs associated to them.

After cross-referencing the datasets, we first examined the properties of the 1,697 control CGIs that were associated with FANTOM2-candidates (FCA-CGIs). No statistically significant differences in the average SINE repeat content (14.2%) or CGI scores (-24.3) of the FCA-CGIs and other controls were observed at the level $p = .05$. This indicates that the FANTOM2-candidate set contains many non-imprinted transcripts. Some of these may be downstream regulatory targets of imprinted genes, as such transcripts will inevitably be included in predictions based on expression profiling. However, the FANTOM2-candidate transcripts are also likely to include novel imprinted genes. We therefore compared the FCA-CGIs with our set of predicted novel imprinted CGIs.

Of the 218 control CGIs that we predicted as novel imprinted loci, 48 (22%) were associated with FANTOM2 candidate imprinted genes. This degree of overlap was significantly higher ($p < 10^{-6}$, X^2 test) than for the 13,401 other

control CGIs, of which only 1,649 (12%) were associated with FANTOM2 candidates. The significant enrichment of FANTOM2-associated CGIs in the set of predicted imprinted loci represents an independent experimental validation of our prediction method and provides further evidence to support the idea that some of our predictions truly are novel DMR-CGIs. However, our computational method appears to be more selective than the FANTOM2 microarray-based experimental approach.

Furthermore, comparison of our set of predicted novel imprinted loci to the independently-determined FANTOM2 candidates offered a unique opportunity to estimate the number of true imprinted genes expected to be present in our predicted set and in the mouse genome overall. By considering the rates at which imprinted and control loci associate with FANTOM2 transcripts, the expected proportion of predicted imprinted loci which are true DMR-CGIs can be calculated in the following way. As 8 of the 27 known DMR-CGIs (30%) and 1,697 of the 13,619 control CGIs (12%) were associated with FANTOM2 transcripts, we may consider these percentages to be estimates for the rates at which all imprinted and control genes will associate with FANTOM2 transcripts. If we then assume that the 218 predicted novel DMR-CGIs contain a mixture of imprinted and control loci, we can determine the proportion of these genes that are expected to be truly imprinted based on the percentage of these loci that are associated with FANTOM2 transcripts. Since 48 of the 218 (22%) predicted novel DMR-CGIs are associated with FANTOM2 transcripts, we would expect that 55% of them to represent true novel imprinted loci (obtained by solving for x in the equation, $0.30x + 0.12[1 - x] = 0.22$). This would suggest that there are 120 novel imprinted DMR-CGIs in our set of 218 predicted candidates. Assuming that the sensitivity rate of our method is the same for novel DMR-CGIs as it is for the 27 known DMR-CGIs (33.3%), we would then expect an additional 240 novel imprinted loci to exist in the entire mouse genome which we failed to predict as DMR-CGIs. Therefore, based on our method we would estimate the total number of DMR-CGIs in the mouse genome to be 387 (including the 27 known DMR-CGIs), which is higher than but not incompatible with previous

estimates placing a lower bound on the number of imprinted loci at 100 (Reik and Walter, 2001).

Although the imprinting status of our predicted DMR-CGIs remains to be determined, this set represents a valuable resource for the analysis of genomic imprinting. The CGIs which are predicted to be novel imprinted loci in both our set and the FANTOM2 set comprise a particularly strong set of imprinting candidates, as they display both sequence properties and expression patterns characteristic of genomic imprinting. The 218 CGIs which we have predicted as novel imprinted loci are fully described in Appendix 2 and this set constitutes a potential resource for focusing experimental identification of new imprinted mouse genes.

DISCUSSION

Genomic sequences from imprinted and control loci in mouse were analyzed according to repeat content and CGI sequence properties in order to identify features of DMR-CGIs, and these features were used to help predict novel imprinted loci. The SINE repeat content at imprinted loci was significantly lower than for controls, demonstrating that this characteristic of imprinted regions which was previously reported in humans is also conserved in mouse. The sequence composition of CGIs associated with imprinted and control genes was examined, and DMR-CGIs were shown to have significantly fewer CpG sites, supporting the hypothesis that differential methylation of imprinted CGIs is reflected in an erosion of their CpG content. A considerable number of oligonucleotides with significantly different frequencies between DMR and control CGIs were also found. Some of these significant oligonucleotides were identified as *CTCF*-binding sites, reflecting the importance of *CTCF* to the process of genomic imprinting, and suggesting a broader role for *CTCF* in the establishment or maintenance of CGIs in general.

A CGI scoring function based on the set of significant oligonucleotides

estimates placing a lower bound on the number of imprinted loci at 100 (Reik and Walter, 2001).

Although the imprinting status of our predicted DMR-CGIs remains to be determined, this set represents a valuable resource for the analysis of genomic imprinting. The CGIs which are predicted to be novel imprinted loci in both our set and the FANTOM2 set comprise a particularly strong set of imprinting candidates, as they display both sequence properties and expression patterns characteristic of genomic imprinting. The 218 CGIs which we have predicted as novel imprinted loci are fully described in Appendix 2 and this set constitutes a potential resource for focusing experimental identification of new imprinted mouse genes.

DISCUSSION

Genomic sequences from imprinted and control loci in mouse were analyzed according to repeat content and CGI sequence properties in order to identify features of DMR-CGIs, and these features were used to help predict novel imprinted loci. The SINE repeat content at imprinted loci was significantly lower than for controls, demonstrating that this characteristic of imprinted regions which was previously reported in humans is also conserved in mouse. The sequence composition of CGIs associated with imprinted and control genes was examined, and DMR-CGIs were shown to have significantly fewer CpG sites, supporting the hypothesis that differential methylation of imprinted CGIs is reflected in an erosion of their CpG content. A considerable number of oligonucleotides with significantly different frequencies between DMR and control CGIs were also found. Some of these significant oligonucleotides were identified as *CTCF*-binding sites, reflecting the importance of *CTCF* to the process of genomic imprinting, and suggesting a broader role for *CTCF* in the establishment or maintenance of CGIs in general.

A CGI scoring function based on the set of significant oligonucleotides

was developed that assigns greater scores to DMR-CGIs than controls at a highly significant level ($p < 10^{-5}$). The CGI scoring function was used in conjunction with regional SINE repeat content to predict novel imprinted loci, and this method yielded a 20-fold enrichment of DMR-CGIs over the original dataset. Genes associated with the predicted novel imprinted loci were compared with another set of candidate imprinted genes identified by large-scale microarray analysis, and a significant overlap between both sets was observed, representing an independent experimental validation of our prediction method.

Examination of the genes associated with the predicted novel imprinted loci revealed interesting trends in their functional annotations. A majority of the 202 associated genes with EnsEMBL-annotated descriptions appear to be involved in pathways related to development and cellular growth, in agreement with the functional characteristics of most known imprinted genes. While some of these genes, such as the *HOX* gene clusters A, B, C, and D, may be unlikely imprinting candidates, the fact that they exhibit similar sequence attributes with imprinted loci suggests possible similarities in the mechanisms of regulation between imprinting and additional pathways. Two genes in particular, the Polycomb Complex Protein *BMI-1* and *RYBP* (Ring1 and YY1 Binding Protein), which were included in our predicted set, are closely associated with the processes of epigenetic regulation, chromatin modification, and developmentally-related gene silencing, suggesting potentially interesting links between these systems of epigenetic regulation that may be explored in future studies. On the other hand, a large proportion (~25%) of the genes associated with predicted imprinted loci are homeobox proteins, and other transcription factors, kinases, phosphatases, and receptor proteins that were also included in our set may represent promising candidate imprinted genes.

In evaluating our method, it is instructive to consider the classification of imprinted genes that were not represented in our intitial dataset of imprinted genes. We therefore examined the performance of our method in classifying 4 genes which were recently discovered to be imprinted (*Gatm*, *DLX5*, *Calcr*, and *A19*) as well as 3 imprinted genes which had been previously known (*U2AF1-rs1*,

Slc38a4, *Peg13*) but were not annotated in the EnsEMBL database (mus_musculus_core_9_3) used at the time of our original analysis.

One of the previously known imprinted genes, *Peg13*, contained a CGI with a high CGI score (56.2) and low SINE content (3.99%) that would clearly have been accurately classified as a DMR-CGI by our method. *U2AF1-rs1* also contained a CGI with a very high CGI score (72.3) but was located in a region of surprisingly high SINE content (20.67%) which precluded its classification as a DMR-CGI. *Slc38a4*, on the other hand, contained a CGI with a SINE content of 7.06% and a CGI score of 14.9, which ranks higher than 90% of all control CGIs but is not sufficient for classification as a DMR-CGI according to our method. The mouse orthologue of the *DLX5* homeobox protein gene, which has recently been shown to be imprinted in humans (Okita *et al*, 2003), was actually included in our set of 218 predicted novel imprinted loci. Although the methylation status of the *DLX5* locus is currently unknown, this result strongly suggests that it is differentially-methylated. Another gene, *Calcr,* which has recently been shown to exhibit tissue-specific imprinting in the brain, was almost also correctly predicted by our method. *Calcr* contained a CGI with a score of 17.4 and SINE content of 4.31% that ranked in the top 3.5% of all control CGIs (471 out of 13619), but was just below our threshold for classification as a DMR-CGI. While the near-classification of *Calcr* and *Slc38a4* as imprinted genes is encouraging, it also suggests a possible refinement of the stringent classification threshold currently used. Another gene, *Gatm*, contained 3 CGIs that all received negative CGI scores and therefore would not be classified as imprinted CGIs by our method; however, this result is consistent with experimental evidence indicating that the *Gatm* locus is not differentially-methylated (Sandell *et al*, 2003). Finally, the *A19* gene was also overlooked by our method because it appears to be regulated by a downstream DMR-CGI associated with *Rasgrf1*, which was already included in our imprinted dataset (de la Puente *et al*, 2002).

These examples illustrate some of the major advantages and limitations of our method for predicting genomic imprinting based on bioinformatic sequence analysis. The fact that *Peg13* and *DLX5* were correctly

identified as imprinted genes represents a strong validation of the success and robustness of our method. Interestingly, *DLX5* was accurately classified as an imprinted gene in our analysis, but was not present in the FANTOM2-candidate imprinted gene set. This example highlights some of the advantages of our sequence-based method for identifying imprinted genes, which avoids many of the limitations inherent to expression profiling-based approaches. Our method is not subject to the availability of samples from specific tissues or developmental stages at which imprinted expression has been established, which could explain why *DLX5* was not contained in the FANTOM2-candidate set. Other constraints of expression-based prediction methods that could be responsible are the required inclusion of probes for not-yet-discovered imprinted genes on the array, as well as technical issues intrinsic to microarray technology, such as errors associated with measurement of low-abundance transcripts. Our prediction method is not limited by these factors, and in principle could be applied to the entire mouse genome without any knowledge of transcripts required *a priori*.

Furthermore, bioinformatic methods such as ours can achieve substantially greater prediction specificity than expression-based methods which inevitably include non-imprinted transcripts that are regulatory targets of imprinted genes. It is also possible that the performance of our method could be further improved through the use of other statistical techniques and classification algorithms, such as support and relevance vector machines (Down and Hubbard, 2002). However, one primary limitation of our method is its inability to identify imprinted genes that are not associated with DMR-CGIs, as exemplified by the case of *Gatm*. Although such cases are in the minority, our prediction method is obviously unsuitable for the identification of this class of imprinted genes. Nevertheless, this study demonstrates that bioinformatic methods do represent a valuable approach for identifying novel imprinted genes that can complement existing experimental strategies and contribute to our knowledge of genomic imprinting.

TABLES & FIGURES

Table 1: Known Imprinted Genes and CpG Islands

Imprinted Gene(s)	EnsEMBL Gene ID	Chromosomal Location	DMR
Nnat	ENSMUSG00000027648	2.158434932-158435545	M
Gnas, Gnasx1, Nesp, Nespas	ENSMUSG00000027523	2.175526919-175528515	P
-	ENSMUSG00000027523	2.175537603-175542938	M
-	ENSMUSG00000027523	2.175569444-175573436	U
Copg2, Copg2as, Copg2as2	ENSMUSG00000025607	6.31047005-31047354	P
-	ENSMUSG00000025607	6.31047674-31047873	P
-	ENSMUSG00000025607	6.31080211-31080466	U
Sgce	ENSMUSG00000004631	6.4453104-4453415	M
-	ENSMUSG00000004631	6.4460514-4460754	M
Nap1-l5	ENSMUSG00000029805	6.59357121-59357436	M
H19	ENSMUSG00000000031	7.133006782-133007370	P
Igf2, Igf2as	ENSMUSG00000000033	7.133086642-133086847	P
-	ENSMUSG00000000033	7.133091687-133097997	P
Mash2/Ascl2	ENSMUSG00000009248	7.133401068-133403152	N
Tapa1/Cd81, Tssc4	ENSMUSG00000037706, ENSMUSG00000037699	7.133489190-133489542	U
Tssc4, Kvlqt1/Kcnq1, Kvlqt1as/Lit1	ENSMUSG00000037699, ENSMUSG00000009545	7.133505316-133505877	P
Kvlqt1/Kcnq1, Kvlqt1as/Lit1	ENSMUSG00000009545	7.133543703-133544475	N
-	ENSMUSG00000009545	7.133732998-133734530	M
-	ENSMUSG00000009545	7.133842243-133842445	N
P57KIP2/Cdkn1c	ENSMUSG00000000154	7.133883008-133883764	N
Slc22a1l/Impt1, P57KIP2/Cdkn1c	ENSMUSG00000037664, ENSMUSG00000000154	7.133896782-133902620	P
Slc22a1l/Impt1, Tssc3/Ip1	ENSMUSG00000037664, ENSMUSG00000010760	7.133943302-133944789	P
Tssc3/Ipl, Nap1-l4/Nap2	ENSMUSG00000010760, ENSMUSG00000010759	7.133990276-133991044	N
Obph1/Osbpl5	ENSMUSG00000037606	7.134182532-134182744	U
Ube3a, Ube3aas	ENSMUSG00000025326	7.48955078-48955908	U

Snrpn, Snurf, Ipw, Pwcr1	ENSMUSG00000000948	7.49379542-49379959	M
Ndn, Magel2	ENSMUSG00000033585, ENSMUSG00000033574	7.51709425-51709854	B
Zfp127/Mkrn3,Zfp127as/Mkrn3as	ENSMUSG00000033564	7.51780860-51781093	M
Frat3, Zfp127	ENSMUSG00000033564, ENSMUSG00000033551	7.51824782-51825751	M
Peg3/Pw1 -	ENSMUSG00000002265 ENSMUSG00000002265	7.6089053-6089383 7.6110648-6112563	U M
Rasgrf1	ENSMUSG00000032356	9.90370716-90371012	P
Zac1, Hymai	ENSMUSG00000019817	10.12974657-12975123	M
Meg1/Grb10 -	ENSMUSG00000020176 ENSMUSG00000020176	11.11953633-11954614 11.11964394-11965591	M N
Dlk/Pref1	ENSMUSG00000040856	12.103716118-103716887	N
Meg3/Gtl2	ENSMUSG00000021268	12.103788454-103788744	P
Dio3	ENSMUSG00000040837	12.104541184-104543033	N
Slc22a3	ENSMUSG00000023828	17.11835678-11835880	U
Igf2r, Igf2ras/Air -	ENSMUSG00000023830 ENSMUSG00000023830	17.12144211-12145609 17.12172251-12173135	M P
Impact -	ENSMUSG00000024423 ENSMUSG00000024423	18.12957829-12958387 18.12989774-12991228	U M
Peg1/Mest	ENSMUSG00000029794	Un.130506604-130506818	M
Asb4	ENSMUSG00000042607	No CGIs	U
Usp29, Usp29as, Zim3	ENSMUSG00000023184	No CGIs	M (Peg3)
Zim1	ENSMUSG00000002266	No CGIs	U
Zfp264	NA	NA	U
Ins2	ENSMUSG00000000215	No CGIs	N
Dcn	ENSMUSG00000019929	No CGIs	U
U2af1-rs1	NA*	NA	M
Htr2a	ENSMUSG00000034997	No CGIs	U
Slc38a4/Ata3	NA*	NA	M
Ins1	ENSMUSG00000035804	No CGIs	U
Peg13	NA*	NA	M

45 CGIs were identified in the gene and 50kb upstream sequences of 41 EnsEMBL gene loci associated with known imprinted mouse genes. Genes are listed here with EnsEMBL Gene IDs and associated CGIs identified by cpgplot (http://www.emboss.org). CGIs are categorized according to the methylated allele (P = paternal, M = maternal, U = unknown, N = Neither, B = both). The 27 CGIs that are differentially-methylated (M or P) are DMR-CGIs, and the remaining 18 (U or N or B) are UMR-CGIs. This annotated database of imprinted genes was generously provided to us by Smith and Kelsey.
* excluded from the original analysis because they were not present in the EnsEMBL mus_musculus_core_9_3 database, but were analyzed subsequently (see Discussion).

Table 2: Number of Significant *k*-mers for UMR & DMR vs Control CGIs

k	α	DMR-CGIs	UMR-CGIs	Expected
5	10^{-2}	294	60	10
5	10^{-3}	154	22	1
5	10^{-4}	92	10	0
5	10^{-5}	62	6	0
6	10^{-2}	580	119	41
6	10^{-3}	234	34	4
6	10^{-4}	128	10	0
6	10^{-5}	72	2	0
7	10^{-2}	1208	372	164
7	10^{-3}	372	68	16
7	10^{-4}	150	16	2
7	10^{-5}	82	4	0

The number of *k*-mers with significantly different frequencies in DMR-CGIs and UMR-CGIs relative to Control CGIs are shown for each combination of word length *k* and significance level α used. In all cases, many significant differences were observed between DMR and Control CGIs. Relatively fewer significant differences were seen between UMR and Control CGIs, indicating that the majority of significant compositional differences between DMR and Control CGIs are related to differential methylation. The number of 'false positives' expected to arise merely by chance based on the significance level and number of *k*-mers tested is shown for comparison.

Figure 1: Clustering Analysis of Significant Heptamers

CLUSTER 1A	LOD
CCGCCGC..	-1.98
.CGCCGCG.	-1.78
GCGCCGC..	-1.62
.CGCCGCC.	-1.92
..GCCGCCG	-2.09
..GCCGCCT	-2.8
CLUSTER 1B	**LOD**
.TGCCGCA	1.89
TTGCCGC.	2.01
.TGCCGCG	1.79

CLUSTER 2	LOD
AAAATTT.	2.67
.AAATTTG	3.26
CAAATTT.	3.26
.AAATTTT	2.67

CLUSTER 3	LOD
.ATGTGAC.	3.25
CATGTGA..	2.93
..TGTCACA	2.72
..TGTGACA	2.72

CLUSTER 4	LOD
..CGGCTCG..	1.41
....GCTCGGC	1.8
...GGCTCGG.	1.57
CTCGGCT....	2.22
.TCGGCTC...	2.02
.TCGGCTG...	1.68

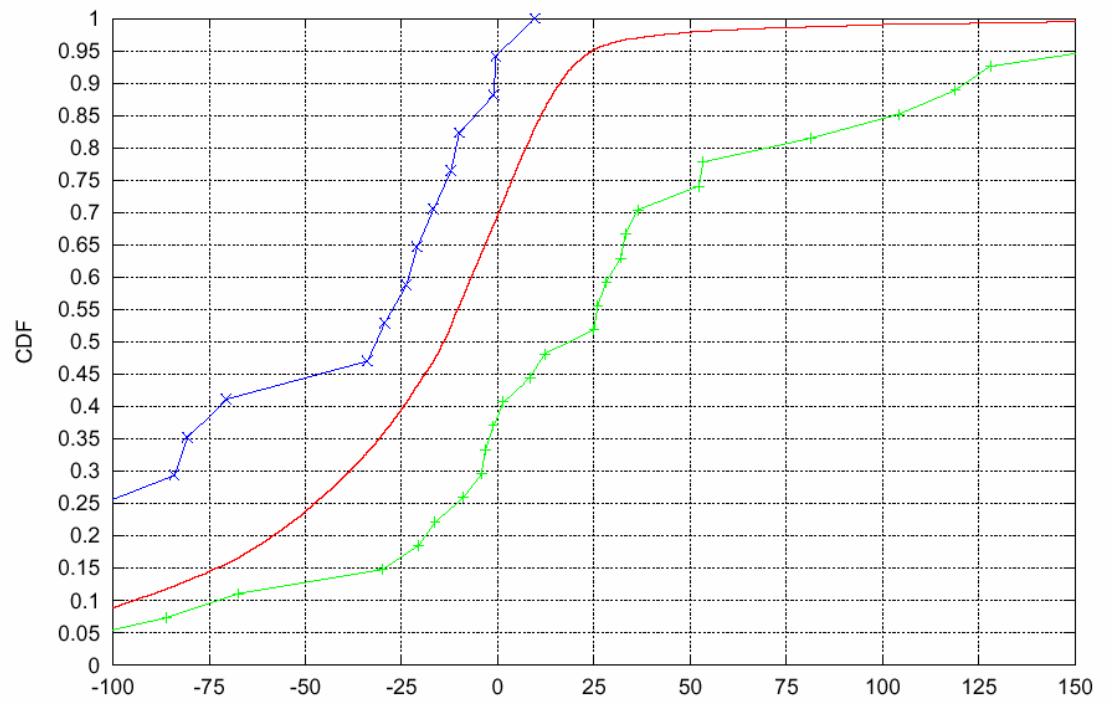
All heptamers present at significantly different ($\alpha = 10^{-5}$) frequencies in DMR and Control CGIs were clustered based on sequence similarity, and four unique clusters with more than three members each were identified. Clustered heptamers and the log-odds ratio in bits (LOD) of their frequencies in DMR and Control CGIs are shown as well as Pictogram representations of motifs for each cluster. Clusters 1A/1B correspond to *CTCF*-binding sites.

Table 3: Results of CGI Scoring Analysis

k	α	Control Avg.	DMR Avg.	>95% Control	*p*-value
5	10^{-2}	-42.4	-24.6	0.19 (5)	8.0E-02
5	10^{-3}	-48.1	-41.2	0.15 (4)	2.6E-01
5	10^{-4}	-43.1	-40.6	0.11 (3)	3.8E-01
5	10^{-5}	-43.1	-40.5	0.07 (2)	3.0E-01
6	10^{-2}	-44.7	-12.3	0.26 (7)	6.0E-04
6	10^{-3}	-36.7	-16.2	0.30 (8)	3.9E-03
6	10^{-4}	-28.6	-10.0	0.26 (7)	7.5E-04
6	10^{-5}	-23.7	-9.6	0.22 (6)	1.7E-03
7	10^{-2}	-26.9	27.8	0.52 (14)	3.7E-06
7	10^{-3}	-16.0	16.6	0.44 (12)	7.0E-07
7	10^{-4}	-8.0	13.5	0.41 (11)	7.5E-05
7	10^{-5}	-5.6	11.0	0.52 (14)	1.6E-05

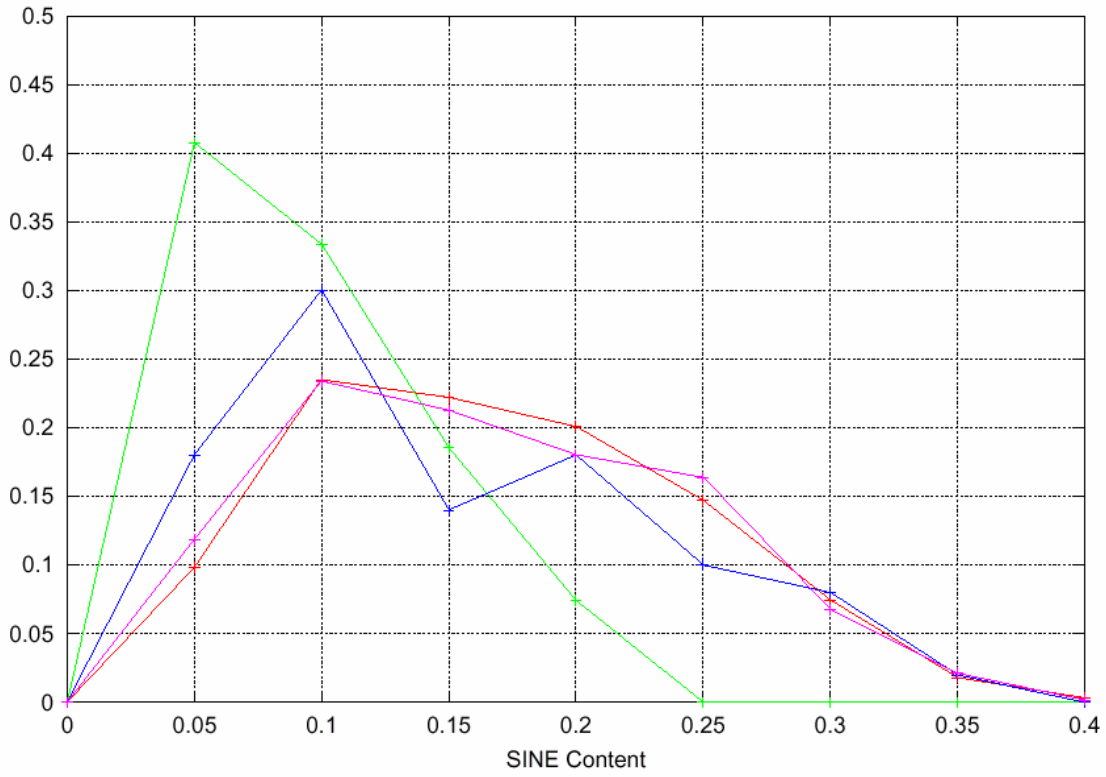
The DMR and Control CGIs were scored for all combinations of word length (k = 5,6,7) and significance level ($\alpha = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$) and many of the score distributions were significantly different for DMR-CGIs and Control CGIs (p-value determined by Wilcoxon rank-sum test). The average CGI Scores for DMR and Control CGIs are shown, along with the proportion of DMR-CGIs that scored higher than 95% of all Control CGIs. The scoring function with parameters (k = 7, $\alpha = 10^{-2}$) was optimal, as it yielded the greatest difference in mean CGI scores between Control and DMR-CGIs and the largest proportion of DMR-CGIs that scored higher than 95% of all Control CGIs.

Figure 2: Cumulative Score Distributions for DMR, UMR & Control CGIs



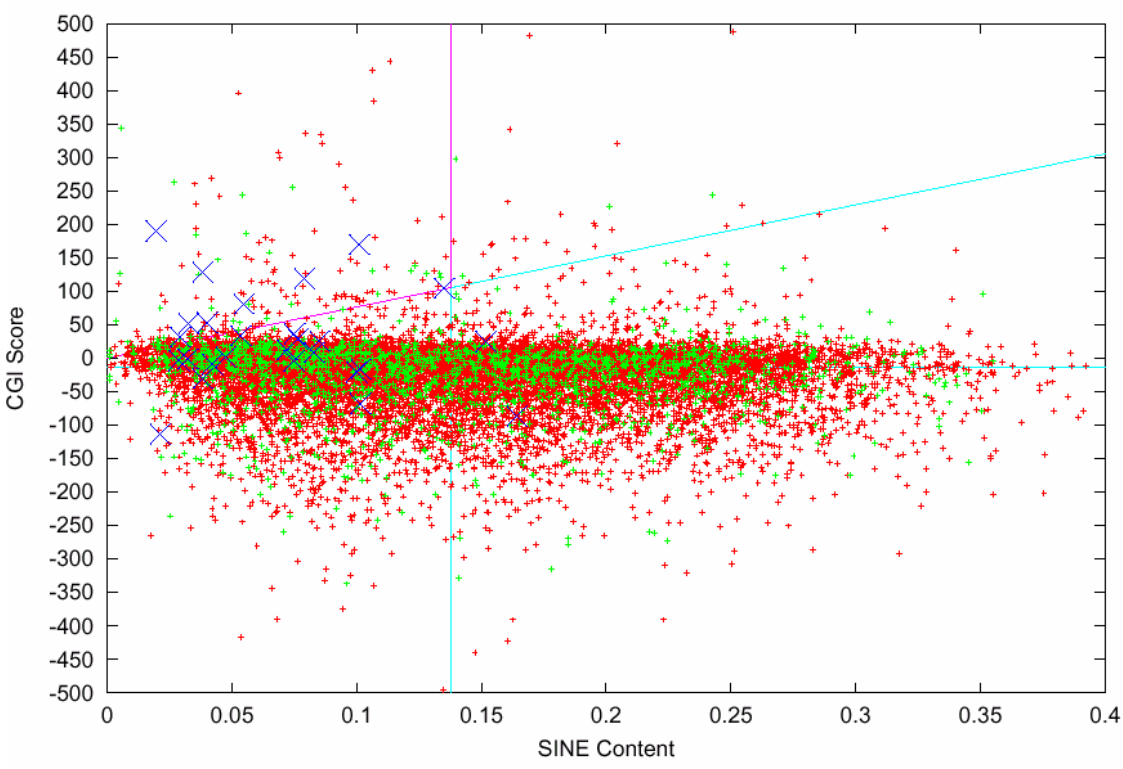
All DMR (green), UMR (blue), and Control (red) CGIs were scored according to their significant heptamer composition ($k = 7$, $\alpha = .01$) as described in methods and the cumulative distribution functions (CDF) were plotted. The average CGI score for DMR-CGIs (27.8) was significantly greater ($p < 10^{-5}$) than for Control CGIs (-26.9), demonstrating that differences in oligonucleotide composition can contribute to the prediction of imprinted loci. The average CGI score for UMR-CGIs (-62.0) was significantly lower than for both DMR-CGIs ($p < 10^{-4}$) and Control CGIs ($p < .05$), indicating that UMR-CGIs are compositionally distinct from DMR-CGIs despite their location in imprinted domains.

Figure 3: SINE Repeat Content Distributions



The SINE repeat content of the sequences and 100kb flanking regions was analyzed for DMR-CGIs (green), Control CGIs (red), High Scoring CGIs (blue) and FANTOM2 Candidate-Associated CGIs (pink). Imprinted loci display a significant reduction in SINE content compared to control regions ($p < 10^{-9}$) that is conserved between mouse and humans. The average SINE content of High Scoring CGIs (12.6%) is significantly greater than for DMR-CGIs (7.4%) and significantly less than for Control CGIs (14.4%), suggesting it contains a mixture of novel imprinted loci together with non-imprinted loci. The mean SINE content of FANTOM2 Candidate-Associated CGIs (14.2%) is not significantly different than for Control CGIs, indicating that many non-imprinted genes (some of which may be downstream regulatory targets of imprinted genes) are included in the FANTOM2-candidate set.

Figure 4: Linear Discriminant Classification of CGIs



A linear discriminant function (LDF) was developed to predict imprinted loci using CGI Scores in conjunction with regional SINE Content. Each CGI is represented as a point in the plane by its SINE Content and CGI Score: DMR-CGIs are blue X's, FCA-CGIs are green dots, and other Control CGIs are red dots. All points that fall above the diagonal pink/aqua line satisfy the LDF [0.012 × CGI Score – 9.175 × SINE Content ≥ 0] and all points to the left of the vertical pink/aqua line have SINE Content lower than the median value for all Control CGIs (0.1378). Points that lie in the upper-left region enclosed by the pink lines meet both criteria and are therefore classified as imprinted loci. 9 DMR-CGIs are correctly classified, while 48 FCA-CGIs and 170 other Control CGIs are predicted to be novel imprinted loci by this method (see Appendix 2).

REFERENCES

Arnaud, P., Monk, D., Hitchins, M., Gordon, E., Dean, W., Beechey, C.V., Peters, J., Craigen, W., Preece, M., Stanier, P., Moore, G.E., and Kelsey, G. (2003). "Conserved methylation imprints in the human and mouse GRB10 genes with divergent allelic expression suggests differential reading of the same mark." Hum Mol Genet 12: 1005-1019.

Down, T.A. and Hubbard, T.J. (2002). "Computational detection and location of transcription start sites in mammalian genomic DNA." Genome Res. 12: 458-461.

Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B. and Bestor, T.H. (2001). "Dnmt3L and the establishment of maternal genomic imprints." Science 294: 2536-2539.

Bird, A.P. (1986). "CpG-rich islands and the function of DNA methylation." Nature 321: 209-213.

Bird, A.P. (2002). "DNA methylation patterns and epigenetic memory." Genes Dev 16: 6-21.

de la Puente, A., Hall, J., Wu, Y.Z., Leone, G., Peters, J., Yoon, B.J., Soloway, P., and Plass, C. (2002). "Structural characterization of Rasgrf1 and a novel linked imprinted locus." Gene 291: 287-297.

Gardiner-Garden, M. and Frommer, M. (1987). "CpG islands in vertebrate genomes." J Mol Biol 196: 261-282.

Greally, J.M. (2002) "Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome." PNAS 99: 327-332.

Ehrlich, M. (2003). “Expression of various genes is controlled by DNA methylation during mammalian development.” J Cell Biochem 88: 899-910.

Hikichi, T., Kohda, T., Kaneko-Ishino, T., and Ishino, F. (2003). “Imprinting regulation of the murine Meg1/Grb10 and human GRB10 genes; roles of brain-specific promoters and mouse-specific CTCF-binding sites.” Nucleic Acids Res 31: 1398-1406.

Ioshikhes, I.P. and Zhang, M.Q. (2000). “Large-scale human promoter mapping using CpG islands.” Nat Genet 26: 61-63.

Ke, X., Thomas, N.S., Robinson, D.O., and Collins, A. (2002) “A novel approach for identifying candidate imprinted genes through sequence analysis of imprinted and control genes.” Hum Genet 111: 511-520.

Kim, J., Kollhoff, A., Bergmann, A, and Stubbs, L. (2003). “Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed gene, Peg3.” Hum Mol Genet 12: 233-245.

Lee, J.T. (2003). “Molecular linkes between X-inactivation and autosomal imprinting: X-inactivation as a driving force for the evolution of imprinting?” Current Biology 13: R242-254.

Li, E., Beard, C. and Jaenisch, R. (1993). “Role for DNA methylation in genomic imprinting.” Nature 366: 362-365.

Nikaido, I., Saito, C., Mizuno, Y., Meguro, M., Bono, H., Kadomura, M., Kono, T., Morris, G.A., Lyons, P.A., Oshimura, M., RIKEN GER Group and GSL Members, Hayashizaki, Y., and Okazaki, Y. (2003). “Discovery of imprinted transcripts in the mouse transcriptome using large-scale expression profiling.” Genome Res 13: 1402-1409.

Ohlsson, R., Renkawitz, R. and Lobanenkov, V. (2001). “CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease.” Trends Genet 17: 520-527.

Okamura, K., Hagiwara-Takeuchi, Y., Li, T., Vu, T.H., Hirai, M., Hattori, M., Sakaki, Y., Hoffman, A.R. and Ito, T. (2000). “Comparative genome analysis of the mouse imprinted gene impact and its nonimprinted human homolog IMPACT: toward the structural basis for species-specific imprinting.” Genome Res 10: 1878-1889.

Okita, C., Meguro, M., Hoshiya, H., Haruta, M., Sakamoto, Y.K., and Oshimura, M. (2003). “A new imprinted cluster on human chromosome 7q21-q31, identified by mouse-human monochromosomal hybrids.” Genomics 81: 556-559.

Paulsen, M., Takada, S., Youngson, N.A., Benchaib, M., Charlier, C., Segers, K., Georges, M., and Ferguson-Smith, A.C. (2001). “Comparative sequence analysis of the imprinted Dlk1-Glt2 locus in three mammalian species reveals highly conserved genomic elements and refines comparison with the Igf2-H19 region.” Genome Res 11: 2085-2094.

Ponger, L., Duret, L., and Mouchiroud, D. (2001). “Determinants of CpG islands: expression in early embryo and isochore structure.” Genome Res 11: 1854-1860.

Reik, W and Walter, J. (2001). “Genomic imprinting: parental influence on the genome.” Nat Rev Genet 2: 21-32.

Sandell, L.L., Guan, X.J., Ingram, R. and Tilghman, S.M. (2003). “Gatm, a creatine synthesis enzyme, is imprinted in mouse placenta.” PNAS 100: 4622-27.

Schoenherr, C.J., Levorse, J.M., and Tilghman, S.M. (2003). “CTCF maintains differential methylation at the Igf2/H19 locus.” Nat Genet 33: 66-69.

APPENDICES

Appendix 1: Frequencies of Significant Heptamers at the level $\alpha = 10^{-5}$

Heptamer	DMR-CGIs	Control CGIs	LOD	*p*-value
ACACACA	0.0002339	0.0006223	-1.41	2.80E-06
AGGCGGC	0.0000439	0.0003064	-2.80	1.39E-06
CACACAC	0.0003070	0.0007220	-1.23	4.30E-06
CCGCCGC	0.0002485	0.0009788	-1.98	3.28E-13
CGCCGCC	0.0002485	0.0009383	-1.92	2.60E-12
CGCCGCG	0.0001316	0.0004528	-1.78	3.52E-06
CGCGGCG	0.0001316	0.0004528	-1.78	3.52E-06
CGGCGGC	0.0001900	0.0008085	-2.09	8.78E-12
GCCGCCG	0.0001900	0.0008085	-2.09	8.78E-12
GCCGCCT	0.0000439	0.0003064	-2.80	1.39E-06
GCGCCGC	0.0001754	0.0005408	-1.62	1.69E-06
GCGGCGC	0.0001754	0.0005408	-1.62	1.69E-06
GCGGCGG	0.0002485	0.0009788	-1.98	3.28E-13
GGCGGCG	0.0002485	0.0009383	-1.92	2.60E-12
GTGTGTG	0.0003070	0.0007220	-1.23	4.30E-06
TGTGTGT	0.0002339	0.0006223	-1.41	2.80E-06
AAAATTT	0.0001462	0.0000230	2.67	8.81E-07
AAATTTG	0.0001608	0.0000168	3.26	3.73E-09
AAATTTT	0.0001462	0.0000230	2.67	8.81E-07
AAGCCCT	0.0002631	0.0000728	1.85	1.33E-06
ACAACTC	0.0001754	0.0000408	2.10	7.64E-06
ACAATGC	0.0001462	0.0000293	2.32	8.45E-06
ACTAATG	0.0001169	0.0000091	3.68	2.25E-08
AGCAGGG	0.0003801	0.0001454	1.39	5.81E-06
AGCCGAG	0.0008332	0.0001789	2.22	3.12E-21
AGGGCTT	0.0002631	0.0000728	1.85	1.33E-06
ATGAGCA	0.0001900	0.0000381	2.32	6.86E-07
ATGTGAC	0.0002339	0.0000246	3.25	3.90E-12
ATTAGTA	0.0001023	0.0000064	4.00	2.22E-08
ATTGCAA	0.0001169	0.0000172	2.77	4.06E-06
ATTTAGG	0.0001316	0.0000216	2.61	3.59E-06
CAAATTT	0.0001608	0.0000168	3.26	3.73E-09
CAACACG	0.0001608	0.0000327	2.30	4.23E-06
CAGCCGA	0.0003070	0.0000959	1.68	1.64E-06
CATGTGA	0.0001900	0.0000249	2.93	4.08E-09
CATTAGT	0.0001169	0.0000091	3.68	2.25E-08
CCCCCAA	0.0003655	0.0001157	1.66	2.90E-07
CCCCCCC	0.0012133	0.0003368	1.85	1.07E-22
CCCTGCT	0.0003801	0.0001454	1.39	5.81E-06

CCGAGCC	0.0008186	0.0002751	1.57	1.08E-12
CCTAAAT	0.0001316	0.0000216	2.61	3.59E-06
CCTGCTC	0.0004970	0.0001958	1.34	6.41E-07
CGAATGA	0.0001169	0.0000173	2.75	4.40E-06
CGAGCCG	0.0004970	0.0001869	1.41	2.26E-07
CGCCAAC	0.0002339	0.0000656	1.83	5.07E-06
CGCGGCA	0.0004532	0.0001309	1.79	1.95E-09
CGGCTCG	0.0004970	0.0001869	1.41	2.26E-07
CGTGTTG	0.0001608	0.0000327	2.30	4.23E-06
CTCGGCT	0.0008332	0.0001789	2.22	3.12E-21
GAGCAGG	0.0004970	0.0001958	1.34	6.41E-07
GAGCCGA	0.0005409	0.0001335	2.02	8.51E-13
GAGTTGT	0.0001754	0.0000408	2.10	7.64E-06
GCATTGT	0.0001462	0.0000293	2.32	8.45E-06
GCCGAGC	0.0009648	0.0002779	1.80	1.00E-17
GCGCCAA	0.0002777	0.0000818	1.76	1.86E-06
GCGGCAA	0.0003508	0.0000871	2.01	5.16E-09
GCTCGGC	0.0009648	0.0002779	1.80	1.00E-17
GGCTCGG	0.0008186	0.0002751	1.57	1.08E-12
GGGGGGG	0.0012133	0.0003368	1.85	1.07E-22
GTCACAT	0.0002339	0.0000246	3.25	3.90E-12
GTGACAA	0.0002777	0.0000349	2.99	1.56E-12
GTTGGCG	0.0002339	0.0000656	1.83	5.07E-06
GTTGTCA	0.0002339	0.0000317	2.88	1.93E-10
TACTAAT	0.0001023	0.0000064	4.00	2.22E-08
TAGTACA	0.0000877	0.0000100	3.13	7.81E-06
TCACATG	0.0001900	0.0000249	2.93	4.08E-09
TCATTCG	0.0001169	0.0000173	2.75	4.40E-06
TCGGCTC	0.0005409	0.0001335	2.02	8.51E-13
TCGGCTG	0.0003070	0.0000959	1.68	1.64E-06
TGACAAC	0.0002339	0.0000317	2.88	1.93E-10
TGCCGCA	0.0003216	0.0000868	1.89	8.35E-08
TGCCGCG	0.0004532	0.0001309	1.79	1.95E-09
TGCGGCA	0.0003216	0.0000868	1.89	8.35E-08
TGCTCAT	0.0001900	0.0000381	2.32	6.86E-07
TGTACTA	0.0000877	0.0000100	3.13	7.81E-06
TGTCACA	0.0002193	0.0000332	2.72	2.80E-09
TGTGACA	0.0002193	0.0000332	2.72	2.80E-09
TTGCAAT	0.0001169	0.0000172	2.77	4.06E-06
TTGCCGC	0.0003508	0.0000871	2.01	5.16E-09
TTGGCGC	0.0002777	0.0000818	1.76	1.86E-06
TTGGGGG	0.0003655	0.0001157	1.66	2.90E-07
TTGTCAC	0.0002777	0.0000349	2.99	1.56E-12

Appendix 2: Predicted Novel DMR-CGIs

The genomic coordinates, CGI Score, regional SINE Content, LDA value, EnsEMBL identifier, and EnsEMBL gene description for all of the 218 predicted novel DMR-CGIs are represented. EnsEMBL gene loci that are associated with FANTOM2 Candidate transcripts are highlighted in green.

#	CGI Locus	Score	SINE	LDA	EnsEMBL Gene ID	Gene Description
1	1.107929988-107930208	13.1	0.0167	0.0043	ENSMUSG00000026327	NA
					ENSMUSG00000026328	18 DAYS EMBRYO CDNA, RIKEN FULL-LENGTH ENRICHED LIBRARY, CLONE=1110013A16, FULL INSERT SEQUENCE.
2	1.139715233-139715939	55.5	0.0474	0.2313	ENSMUSG00000019230	LIM/HOMEOBOX PROTEIN LHX9.
3	1.140669281-140669512	24.1	0.0082	0.2140	ENSMUSG00000026362	COMPLEMENT FACTOR H-RELATED PROTEIN.
4	1.140750137-140750383	8.1	0.0089	0.0165	ENSMUSG00000033898	COMPLEMENT FACTOR H-RELATED PROTEIN.
5	1.145642591-145642940	23.1	0.0260	0.0390	ENSMUSG00000026357	REGULATOR OF G-PROTEIN SIGNALING 18 (RGS18).
6	1.167571950-167572874	91.7	0.0719	0.4435	ENSMUSG00000026563	HYPOTHETICAL 37.4 KDA PROTEIN.
7	1.168898960-168903083	180.0	0.0636	1.5803	ENSMUSG00000026686	LIM HOMEODOMAIN-CONTAINING TRANSCRIPTION FACTOR.
8	1.191520093-191526544	113.4	0.0617	0.7975	ENSMUSG00000010175	HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1).
9	1.34631926-34633988	59.0	0.0641	0.1208	ENSMUSG00000026131	BULLOUS PEMPHIGOID ANTIGEN 1-B.
10	1.44310462-44310729	40.7	0.0444	0.0818	ENSMUSG00000026052	NA
11	1.44313593-44313893	40.4	0.0417	0.1036	ENSMUSG00000026052	NA
12	1.4500872-4506849	255.4	0.0745	2.3870	ENSMUSG00000025902	TRANSCRIPTION FACTOR SOX-17.
13	1.5575123-5575413	20.2	0.0196	0.0629	ENSMUSG00000025905	KAPPA-TYPE OPIOID RECEPTOR (KOR-1) (MSL-1).
14	1.76346798-76349491	92.9	0.0979	0.2184	ENSMUSG00000006576	ANION EXCHANGE PROTEIN 3 (NEURONAL BAND 3-LIKE PROTEIN).
					ENSMUSG00000026213	NA
					ENSMUSG00000032968	INHIBIN ALPHA CHAIN PRECURSOR.
15	2.110630098-110632418	49.9	0.0464	0.1734	ENSMUSG00000041693	BRAIN-DERIVED NEUROTROPHIC FACTOR PRECURSOR (BDNF).
16	2.110646455-110648177	37.6	0.0487	0.0054	ENSMUSG00000041693	BRAIN-DERIVED NEUROTROPHIC FACTOR PRECURSOR (BDNF).
17	2.148392152-148395933	143.6	0.0638	1.1409	ENSMUSG00000037034	PAIRED BOX PROTEIN PAX-1.
18	2.149075615-149080794	175.8	0.0661	1.5063	ENSMUSG00000037025	HEPATOCYTE NUCLEAR FACTOR 3-BETA (HNF-3B).
19	2.149441396-149442253	48.1	0.0582	0.0439	ENSMUSG00000027437	THROMBOMODULIN PRECURSOR (FETOMODULIN) (TM).
20	2.162898605-162901835	158.3	0.0652	1.3042	ENSMUSG00000027413	PROTEIN TYROSINE PHOSPHATASE, RECEPTOR TYPE, T.
21	2.165117351-165120836	78.8	0.0877	0.1429	ENSMUSG00000040164	K+ VOLTAGE-GATED CHANNEL, SUBFAMILY S, 1.
22	2.18817259-18822395	189.8	0.0831	1.5189	ENSMUSG00000026739	POLYCOMB COMPLEX PROTEIN BMI-1.
23	2.25854939-25860390	121.7	0.1335	0.2388	ENSMUSG00000026942	TNF RECEPTOR ASSOCIATED FACTOR 2 (TRAF2).

					ENSMUSG00000036504	NA
24	2.55961454-55962223	42.2	0.0382	0.1562	ENSMUSG00000026824	G PROTEIN-ACTIVATED INWARD RECTIFIER POTASSIUM CHANNEL 1 (GIRK1) (POTASSIUM CHANNEL, INWARDLY RECTIFYING, SUBFAMILY J, MEMBER 3) (INWARD RECTIFIER K+ CHANNEL KIR3.1) (KGA) (KGB1).
25	2.57640962-57642961	101.9	0.0699	0.5836	ENSMUSG00000026826	ORPHAN NUCLEAR RECEPTOR NURR1 (NUR-RELATED FACTOR 1).
26	2.72320504-72325498	130.5	0.1218	0.4514	ENSMUSG00000041911	HOMEOBOX PROTEIN DLX-1.
27	2.74106634-74116088	442.8	0.1133	4.2829	ENSMUSG00000008226	NA
28	2.75516613-75530035	613.6	0.0701	6.7331	ENSMUSG00000001817	HOMEOBOX PROTEIN HOX-D10 (HOX-4.5) (HOX-5.3).
					ENSMUSG00000001819	HOMEOBOX PROTEIN HOX-D13 (HOX-4.8).
					ENSMUSG00000001823	HOMEOBOX PROTEIN HOX-D12 (HOX-4.7) (HOX-5.6).
					ENSMUSG00000042482	HOMEOBOX PROTEIN HOX-D9 (HOX-4.4) (HOX-5.2).
					ENSMUSG00000042499	HOMEOBOX PROTEIN HOX-D11 (HOX-4.6) (HOX-5.5).
29	2.75543059-75555527	644.5	0.0365	7.4138	ENSMUSG00000001815	HOMEOBOX EVEN-SKIPPED HOMOLOG PROTEIN 2 (EVX-2).
					ENSMUSG00000027102	HOMEOBOX PROTEIN HOX-D8 (HOX-4.3) (HOX-5.4).
					ENSMUSG00000042464	HOMEOBOX PROTEIN HOX-D4 (HOX-4.2) (HOX-5.1).
					ENSMUSG00000042482	HOMEOBOX PROTEIN HOX-D9 (HOX-4.4) (HOX-5.2).
30	2.75565849-75573593	83.9	0.0230	0.7972	ENSMUSG00000042453	HOMEOBOX PROTEIN HOX-D3 (HOX-4.1) (MH-19).
31	2.80334667-80337688	194.7	0.0355	2.0146	ENSMUSG00000034701	NEUROGENIC DIFFERENTIATION FACTOR 1 (NEUROD1).
32	2.84864734-84865841	47.4	0.0394	0.2088	ENSMUSG00000034552	NA
33	2.94622704-94627599	110.6	0.0672	0.7132	ENSMUSG00000040310	HOMEOBOX PROTEIN ARISTALESS-LIKE 4 (ALX-4).
34	2.9912809-9918564	211.8	0.1344	1.3131	ENSMUSG00000015619	TRANS-ACTING T-CELL SPECIFIC TRANSCRIPTION FACTOR GATA-3.
					ENSMUSG00000025783	NA
35	3.129708234-129713310	173.3	0.0610	1.5237	ENSMUSG00000028023	PITUITARY HOMEOBOX 2 (ORTHODENTICLE-LIKE HOMEOBOX 2) (SOLURSHIN) (ALL1 RESPONSIVE PROTEIN ARP1) (BRX1 HOMEOPROTEIN) (PAIRED-LIKE HOMEODOMAIN TRANSCRIPTION FACTOR MUNC 30).
36	3.131661313-131667880	96.6	0.1087	0.1640	ENSMUSG00000027985	LYMPHOID ENHANCER BINDING FACTOR 1 (LEF-1).
37	3.154910223-154915320	308.0	0.0688	3.0720	ENSMUSG00000028201	LIM/HOMEOBOX PROTEIN LHX8 (L3).
38	3.17768553-17771839	35.8	0.0335	0.1223	ENSMUSG00000039527	BASIC HELIX-LOOP-HELIX DOMAIN CONTAINING, CLASS B5
39	3.29825267-29826282	53.1	0.0388	0.2819	ENSMUSG00000027684	ECOTROPIC VIRUS INTEGRATION 1 SITE PROTEIN.
40	3.45477661-45482475	184.8	0.0355	1.8952	ENSMUSG00000027730	NA
					ENSMUSG00000037927	PROTOCADHERIN 10
41	3.55275165-55278987	148.6	0.0974	0.8924	ENSMUSG00000036615	REGULATORY FACTOR X-ASSOCIATED PROTEIN.
42	3.60382842-60383258	24.1	0.0056	0.2384	ENSMUSG00000036976	NA
43	3.67580670-67584119	79.9	0.0740	0.2814	ENSMUSG00000027833	SHORT STATURE HOMEOBOX PROTEIN 2 (HOMEOBOX PROTEIN OG12X) (OG-12) (PAIRED FAMILY HOMEODOMAIN PROTEIN PRX3).
					ENSMUSG00000034544	NA

44	3.96544118-96552984	384.2	0.1066	3.6401	ENSMUSG00000038445	HISTONE H3 (H3.2).
45	3.96791854-96794377	100.1	0.1132	0.1640	ENSMUSG00000038403	NA
46	4.11571701-11574111	147.0	0.0577	1.2375	ENSMUSG00000040642	NA
47	4.43550588-43552290	117.2	0.1264	0.2500	ENSMUSG00000014030	PAIRED BOX PROTEIN PAX-5 (B-CELL SPECIFIC TRANSCRIPTION FACTOR) (BSAP).
48	4.51921727-51922560	65.4	0.0438	0.3836	ENSMUSG00000015243	ATP-BINDING CASSETTE, SUB-FAMILY A, MEMBER 1 (ATP-BINDING CASSETTE TRANSPORTER 1) (ATP-BINDING CASSETTE 1) (ABC-1).
49	4.79784817-79794020	243.5	0.0544	2.4289	ENSMUSG00000008575	NUCLEAR FACTOR 1 B-TYPE (NUCLEAR FACTOR 1/B) (NF1-B) (NFI-B) (NF-I/B) (CCAAT-BOX BINDING TRANSCRIPTION FACTOR) (CTF) (TGGCA-BINDING PROTEIN).
50	4.93781858-93782066	22.1	0.0268	0.0190	ENSMUSG00000028571	SIMILAR TO CYP2J4.
51	5.105394792-105400633	143.0	0.1236	0.5851	ENSMUSG00000029275	ZINC FINGER PROTEIN GFI-1 (GROWTH FACTOR INDEPENDENCE-1).
52	5.127399336-127400484	88.3	0.0913	0.2243	ENSMUSG00000034268	40S RIBOSOMAL PROTEIN S16.
53	5.18627019-18627391	28.6	0.0246	0.1184	ENSMUSG00000040003	ACTIVIN RECEPTOR INTERACTING PROTEIN 1.
54	5.20644740-20645372	30.7	0.0372	0.0279	ENSMUSG00000038490	REELIN PRECURSOR (EC 3.4.21.-) (REELER PROTEIN).
55	5.34282539-34284302	115.4	0.0576	0.8581	ENSMUSG00000029097	NA
					ENSMUSG00000029098	ACYL-COENZYME A OXIDASE 3, PEROXISOMAL (EC 1.3.3.6) (PRISTANOYL-COA OXIDASE).
56	5.36724455-36728222	107.3	0.0783	0.5719	ENSMUSG00000039404	HOMEOBOX PROTEIN MSX-1 (HOX-7) (HOX-7.1).
57	5.40734938-40740238	268.3	0.0420	2.8401	ENSMUSG00000029129	HOMEOBOX PROTEIN NKX-3.2 (BAGPIPE HOMEOBOX PROTEIN HOMOLOG 1).
58	5.70959108-70959357	21.8	0.0228	0.0532	ENSMUSG00000029212	GAMMA-AMINOBUTYRIC-ACID RECEPTOR BETA-1 SUBUNIT PRECURSOR (GABA(A) RECEPTOR).
59	5.71184005-71184388	27.9	0.0269	0.0879	ENSMUSG00000029212	GAMMA-AMINOBUTYRIC-ACID RECEPTOR BETA-1 SUBUNIT PRECURSOR (GABA(A) RECEPTOR).
60	5.95527555-95528188	36.0	0.0378	0.0860	ENSMUSG00000029338	HYPOTHETICAL 24.5 KDA PROTEIN (FRAGMENT).
61	6.127436907-127437245	39.7	0.0465	0.0512	ENSMUSG00000038097	VOLTAGE-GATED POTASSIUM CHANNEL PROTEIN KV1.5 (KV1-5).
62	6.127540710-127542681	51.5	0.0620	0.0502	ENSMUSG00000003015	VOLTAGE-GATED POTASSIUM CHANNEL PROTEIN KV1.1 (MK1) (MBK1).
63	6.129836345-129836967	29.2	0.0291	0.0843	ENSMUSG00000000248	NA
64	6.130873062-130873277	24.3	0.0139	0.1651	ENSMUSG00000030173	KILLER CELL LECTIN-LIKE RECEPTOR 5 (T-CELL SURFACE GLYCOPROTEIN LY-49E) (LY49-E ANTIGEN).
65	6.17642460-17642693	42.4	0.0527	0.0263	ENSMUSG00000029534	SUPPRESSION OF TUMORIGENICITY 7
66	6.18782659-18784598	30.3	0.0328	0.0633	ENSMUSG00000029517	NA
67	6.23207766-23211246	152.8	0.0743	1.1548	ENSMUSG00000029697	NA
68	6.52593717-52596796	112.1	0.0045	1.3065	ENSMUSG00000029844	HOMEOBOX PROTEIN HOX-A1 (HOX-1.6) (HOMEOTIC PROTEIN ERA-1-993) (EARLY RETINOIC ACID 1) (HOMEOBOXLESS PROTEIN ERA-1-399).
69	6.52596553-52596796	4.8	0.0038	0.0221	ENSMUSG00000038243	HOMEOBOX PROTEIN HOX-A6 (HOX-1.2) (M5-4).
70	6.52601791-52603517	56.4	0.0038	0.6439	ENSMUSG00000000942	HOMEOBOX PROTEIN HOX-A4 (HOX-1.4) (MH-3).

					ENSMUSG00000014704	HOMEOBOX PROTEIN HOX-A2 (HOX-1.11).
					ENSMUSG00000038243	HOMEOBOX PROTEIN HOX-A6 (HOX-1.2) (M5-4).
					ENSMUSG00000038253	HOMEOBOX PROTEIN HOX-A5 (HOX-1.3) (M2).
					ENSMUSG00000038270	HOMEOBOX PROTEIN HOX-A3 (HOX-1.5) (MO-10).
71	6.52606992-52612049	343.9	0.0055	4.0843	ENSMUSG00000000942	HOMEOBOX PROTEIN HOX-A4 (HOX-1.4) (MH-3).
					ENSMUSG00000038236	HOMEOBOX PROTEIN HOX-A7 (HOX-1.1) (M6-12) (M6).
					ENSMUSG00000038243	HOMEOBOX PROTEIN HOX-A6 (HOX-1.2) (M5-4).
					ENSMUSG00000038253	HOMEOBOX PROTEIN HOX-A5 (HOX-1.3) (M2).
					ENSMUSG00000038270	HOMEOBOX PROTEIN HOX-A3 (HOX-1.5) (MO-10).
72	6.52614853-52615076	8.2	0.0107	0.0008	ENSMUSG00000000942	HOMEOBOX PROTEIN HOX-A4 (HOX-1.4) (MH-3).
					ENSMUSG00000038227	HOMEOBOX PROTEIN HOX-A9 (HOX-1.7).
					ENSMUSG00000038236	HOMEOBOX PROTEIN HOX-A7 (HOX-1.1) (M6-12) (M6).
					ENSMUSG00000038243	HOMEOBOX PROTEIN HOX-A6 (HOX-1.2) (M5-4).
					ENSMUSG00000038253	HOMEOBOX PROTEIN HOX-A5 (HOX-1.3) (M2).
73	6.52615998-52621268	262.3	0.0270	2.9054	ENSMUSG00000000942	HOMEOBOX PROTEIN HOX-A4 (HOX-1.4) (MH-3).
					ENSMUSG00000038227	HOMEOBOX PROTEIN HOX-A9 (HOX-1.7).
					ENSMUSG00000038236	HOMEOBOX PROTEIN HOX-A7 (HOX-1.1) (M6-12) (M6).
					ENSMUSG00000038243	HOMEOBOX PROTEIN HOX-A6 (HOX-1.2) (M5-4).
					ENSMUSG00000038253	HOMEOBOX PROTEIN HOX-A5 (HOX-1.3) (M2).
74	6.61427657-61427934	41.3	0.0212	0.3026	ENSMUSG00000025891	NA
75	6.61465369-61465636	21.4	0.0253	0.0249	ENSMUSG00000025891	NA
76	6.6626082-6628712	129.1	0.0668	0.9397	ENSMUSG00000029755	HOMEOBOX PROTEIN DLX-5.
77	6.68701932-68702134	11.1	0.0130	0.0142	ENSMUSG00000029895	IG KAPPA CHAIN V-II REGION VKAPPA167 PRECURSOR.
78	6.78137598-78139829	86.7	0.1026	0.1016	ENSMUSG00000030026	ALPHA-2 CATENIN (ALPHA-CATENIN RELATED PROTEIN) (ALPHA N-CATENIN).
					ENSMUSG00000037682	NA
79	6.89060972-89067216	180.9	0.1074	1.1894	ENSMUSG00000015053	ENDOTHELIAL TRANSCRIPTION FACTOR GATA-2.
80	6.90977417-90978167	23.5	0.0228	0.0730	ENSMUSG00000034468	VN5 (VOMERONASAL RECEPTOR V1RB3).
81	7.127391482-127405249	396.4	0.0527	4.2814	ENSMUSG00000010476	TRANSCRIPTION FACTOR COE3 (EARLY B-CELL FACTOR 3) (EBF-3) (OLF-1/EBF- LIKE 2) (OE-2) (O/E-2).
82	7.13599839-13600074	10.8	0.0121	0.0196	ENSMUSG00000005602	MYOTUBULARIN-RELATED PROTEIN 2 (FRAGMENT).
					ENSMUSG00000008991	CEA13 PROTEIN (FRAGMENT).
83	7.19240894-19241512	46.1	0.0242	0.3316	ENSMUSG00000003017	CYTOCHROME P450 2A12 (EC 1.14.14.1) (CYPIIA12) (STEROID HORMONES 7- ALPHA-HYDROXYLASE) (TESTOSTERONE 7-ALPHA-HYDROXYLASE).
84	7.38114850-38115524	22.8	0.0185	0.1045	ENSMUSG00000030476	G PROTEIN-COUPLED RECEPTOR.
85	7.39256410-39259340	205.4	0.1246	1.3264	ENSMUSG00000030507	HOMEOBOX PROTEIN DBX.
86	7.46828784-46828999	33.8	0.0329	0.1046	ENSMUSG00000030449	GAMMA-AMINOBUTYRIC-ACID RECEPTOR GAMMA-3 SUBUNIT PRECURSOR (GABA(A) RECEPTOR).
87	7.6640131-6640338	6.8	0.0075	0.0127	ENSMUSG00000000605	CHLORIDE CHANNEL PROTEIN 4 (CLC-4).

					ENSMUSG00000034155	ADULT MALE TESTIS CDNA, RIKEN FULL-LENGTH ENRICHED LIBRARY, CLONE=4930547K11, FULL INSERT SEQUENCE.
88	7.94712714-94712935	14.7	0.0141	0.0481	ENSMUSG00000041946	MITOGEN-ACTIVATED PROTEIN KINASE KINASE 1 INTERACTING PROTEIN 1
89	8.100420881-100421409	13.2	0.0156	0.0159	ENSMUSG00000035880	RNA FOR TYPE IIB INTRACISTERNAL A-PARTICLE (IAP) ELEMENT ENCODING INTEGRASE, CLONE 106 (IAP) (RNA FOR TYPE IIB INTRACISTERNAL A-PARTICLE (IAP) ELEMENT ENCODING INTEGRASE, CLONE 111).
90	8.100424459-100424746	17.3	0.0170	0.0527	ENSMUSG00000035880	RNA FOR TYPE IIB INTRACISTERNAL A-PARTICLE (IAP) ELEMENT ENCODING INTEGRASE, CLONE 106 (IAP) (RNA FOR TYPE IIB INTRACISTERNAL A-PARTICLE (IAP) ELEMENT ENCODING INTEGRASE, CLONE 111).
91	8.104629960-104630191	26.1	0.0252	0.0824	ENSMUSG00000031884	SIMILAR TO CARBOXYLESTERASE 2 (INTESTINE, LIVER).
92	8.104661963-104662473	19.8	0.0196	0.0585	ENSMUSG00000031884	SIMILAR TO CARBOXYLESTERASE 2 (INTESTINE, LIVER).
93	8.128411535-128414389	97.9	0.1151	0.1211	ENSMUSG00000025810	NEUROPILIN-1 PRECURSOR (A5 PROTEIN).
94	8.21535234-21538161	142.4	0.1003	0.7916	ENSMUSG00000031539	ADAPTOR-RELATED PROTEIN COMPLEX AP-3 MU2 SUBUNIT.
95	8.44172592-44173463	80.1	0.0570	0.4398	ENSMUSG00000031646	MOUSE FAT 1 CADHERIN (FRAGMENT).
96	8.44186073-44189712	186.8	0.0559	1.7321	ENSMUSG00000031646	MOUSE FAT 1 CADHERIN (FRAGMENT).
					ENSMUSG00000038952	MELATONIN RECEPTOR TYPE 1A (MEL-1A-R).
97	8.44498250-44499005	41.1	0.0498	0.0373	ENSMUSG00000031640	PLASMA KALLIKREIN PRECURSOR (EC 3.4.21.34) (PLASMA PREKALLIKREIN) (KININOGENIN) (FLETCHER FACTOR).
98	8.91555744-91561784	43.0	0.0559	0.0041	ENSMUSG00000031734	IROQUOIS-CLASS HOMEODOMAIN PROTEIN IRX-3.
99	8.9936844-9938878	74.1	0.0416	0.5093	ENSMUSG00000031497	TUMOR NECROSIS FACTOR LIGAND SUPERFAMILY MEMBER 13B (B CELL-ACTIVATING FACTOR) (BAFF).
					ENSMUSG00000040396	NA
100	9.38502882-38503084	14.2	0.0157	0.0258	ENSMUSG00000040248	PUTATIVE OLFACTORY RECEPTOR (FRAGMENT).
101	9.56065428-56072433	236.8	0.0984	1.9437	ENSMUSG00000032314	SIMILAR TO ELECTRON-TRANSFER-FLAVOPROTEIN, ALPHA POLYPEPTIDE (GLUTARIC ACIDURIA II).
102	9.75504124-75505553	90.4	0.1004	0.1658	ENSMUSG00000034924	HEPATOCYTE NUCLEAR FACTOR 6 (HNF-6) (ONE CUT DOMAIN FAMILY MEMBER 1).
103	9.87409038-87412203	94.3	0.0918	0.2909	ENSMUSG00000032415	NA
104	9.91848564-91852357	139.2	0.0297	1.4004	ENSMUSG00000032368	ZINC FINGER PROTEIN ZIC1 (ZINC FINGER PROTEIN OF THE CEREBELLUM 1).
					ENSMUSG00000036972	ZINC FINGER PROTEIN ZIC4 (ZINC FINGER PROTEIN OF THE CEREBELLUM 4).
105	9.91867752-91870711	155.4	0.0368	1.5305	ENSMUSG00000036972	ZINC FINGER PROTEIN ZIC4 (ZINC FINGER PROTEIN OF THE CEREBELLUM 4).
106	10.103417472-103417775	45.0	0.0474	0.1053	ENSMUSG00000019892	NA
107	10.121627379-121630233	152.2	0.1359	0.5833	ENSMUSG00000034707	NA
108	10.128394146-128398404	67.5	0.0641	0.2238	ENSMUSG00000025399	RETINOL DEHYDROGENASE TYPE 6.
109	10.25395244-25395985	242.2	0.0451	2.4982	ENSMUSG00000019978	BAND 4.1-LIKE PROTEIN 2 (GENERALLY EXPRESSED PROTEIN 4.1) (4.1G).

110	10.35743727-35743941	17.3	0.0199	0.0252	ENSMUSG00000039439	S-ADENOSYLMETHIONINE DECARBOXYLASE PROENZYME 1 (EC 4.1.1.50) (ADOMETDC 1) (SAMDC 1)
111	10.41525848-41527854	99.9	0.1062	0.2267	ENSMUSG00000019821	NA
112	11.11316328-11316706	22.1	0.0240	0.0451	ENSMUSG00000020193	ZONA PELLUCIDA BINDING PROTEIN.
113	11.21593389-21594262	35.5	0.0433	0.0296	ENSMUSG00000020321	MALATE DEHYDROGENASE, CYTOPLASMIC (EC 1.1.1.37).
114	11.5889970-5892641	98.1	0.0538	0.6851	ENSMUSG00000020465	CALMODULIN-DEPENDENT PROTEIN KINASE II BETA M ISOFORM (FRAGMENT).
					ENSMUSG00000020466	CALCIUM/CALMODULIN-DEPENDENT PROTEIN KINASE TYPE II BETA CHAIN (CAM- KINASE II BETA CHAIN) (EC 2.7.1.123) (CAMK-II, BETA SUBUNIT).
115	11.59486736-59487762	47.8	0.0596	0.0273	ENSMUSG00000020455	TRIPARTITE MOTIF PROTEIN TRIM11.
					ENSMUSG00000020496	UNKNOWN PROTEIN (FRAGMENT).
					ENSMUSG00000036952	TRIPARTITE MOTIF PROTEIN TRIM17 (FRAGMENT).
116	11.97047828-97048047	20.9	0.0262	0.0107	ENSMUSG00000000690	HOMEOBOX PROTEIN HOX-B6 (HOX-2.2) (MH-22A).
					ENSMUSG00000038684	HOMEOBOX PROTEIN HOX-B3 (HOX-2.7) (MH-23).
					ENSMUSG00000038692	HOMEOBOX PROTEIN HOX-B4 (HOX-2.6).
					ENSMUSG00000038700	HOMEOBOX PROTEIN HOX-B5 (HOX-2.1) (MU-1) (H24.1).
117	11.97052217-97054059	17.5	0.0101	0.1179	ENSMUSG00000000690	HOMEOBOX PROTEIN HOX-B6 (HOX-2.2) (MH-22A).
					ENSMUSG00000038684	HOMEOBOX PROTEIN HOX-B3 (HOX-2.7) (MH-23).
					ENSMUSG00000038692	HOMEOBOX PROTEIN HOX-B4 (HOX-2.6).
					ENSMUSG00000038700	HOMEOBOX PROTEIN HOX-B5 (HOX-2.1) (MU-1) (H24.1).
118	11.97056320-97058384	127.1	0.0050	1.4812	ENSMUSG00000038684	HOMEOBOX PROTEIN HOX-B3 (HOX-2.7) (MH-23).
					ENSMUSG00000038692	HOMEOBOX PROTEIN HOX-B4 (HOX-2.6).
					ENSMUSG00000038700	HOMEOBOX PROTEIN HOX-B5 (HOX-2.1) (MU-1) (H24.1).
119	12.108799297-108799519	15.6	0.0123	0.0753	ENSMUSG00000003771	IG MU CHAIN C REGION.
120	12.108908897-108909473	10.8	0.0015	0.1165	ENSMUSG00000003771	IG MU CHAIN C REGION.
					ENSMUSG00000037192	IG HEAVY CHAIN V REGION 36-65.
121	12.109727905-109728117	14.7	0.0084	0.0991	ENSMUSG00000002716	IG HEAVY CHAIN V REGION 23 PRECURSOR.
122	12.109728895-109729173	8.1	0.0084	0.0203	ENSMUSG00000002716	IG HEAVY CHAIN V REGION 23 PRECURSOR.
123	12.51155681-51158627	151.5	0.0559	1.3091	ENSMUSG00000001497	PAIRED BOX PROTEIN PAX-9.
124	12.52685216-52685629	36.9	0.0328	0.1427	ENSMUSG00000035431	SOMATOSTATIN RECEPTOR TYPE 1 (SS1R) (SRIF-2).
125	13.115198010-115198363	15.4	0.0199	0.0020	ENSMUSG00000021730	HYPERPOLARIZATION-ACTIVATED, CYCLIC NUCLEOTIDE-GATED K+ 1.
126	13.21140069-21142250	66.4	0.0509	0.3309	ENSMUSG00000008648	HISTONE H3.1 (H3/A) (H3/C) (H3/D) (H3/F) (H3/H) (H3/I) (H3/J) (H3/K) (H3/L).
					ENSMUSG00000016909	HISTONE H2B 291B.
					ENSMUSG00000016977	HISTONE H2B F (H2B 291A).
					ENSMUSG00000036577	HISTONE H4.
127	13.21467693-21470325	131.9	0.0669	0.9722	ENSMUSG00000006179	THYMUS-SPECIFIC SERINE PROTEASE PRECURSOR (EC 3.4.-.-).
128	13.22831204-22833551	115.4	0.0645	0.7960	ENSMUSG00000036376	ACTIVATOR OF BASAL TRANSCRIPTION.

129	13.22943293-22945160	91.8	0.0954	0.2273	ENSMUSG00000018084	HISTONE H3.1 (H3/A) (H3/C) (H3/D) (H3/F) (H3/H) (H3/I) (H3/J) (H3/K) (H3/L).
					ENSMUSG00000018094	HISTONE H2B F (H2B 291A).
					ENSMUSG00000018100	HISTONE H1.3 (H1 VAR.4) (H1D).
					ENSMUSG00000018101	HISTONE H2A.G (H2A/G) (H2A.3).
					ENSMUSG00000036243	HISTONE H4.
					ENSMUSG00000036253	HISTONE H3 (H3.2).
					ENSMUSG00000036326	HISTONE H3 (H3.2).
130	13.23154492-23156486	51.3	0.0608	0.0590	ENSMUSG00000006611	HEREDITARY HAEMOCHROMATOSIS PROTEIN HOMOLOG PRECURSOR.
					ENSMUSG00000016575	HISTONE H2A.G (H2A/G) (H2A.3).
					ENSMUSG00000036132	HISTONE H1.1 (H1 VAR.3) (H1A).
					ENSMUSG00000036149	HISTONE H3 (H3.2).
					ENSMUSG00000036173	HISTONE H3 (H3.2).
					ENSMUSG00000036201	HISTONE H4.
131	13.23160436-23162013	86.1	0.0585	0.4976	ENSMUSG00000006611	HEREDITARY HAEMOCHROMATOSIS PROTEIN HOMOLOG PRECURSOR.
					ENSMUSG00000036132	HISTONE H1.1 (H1 VAR.3) (H1A).
					ENSMUSG00000036149	HISTONE H3 (H3.2).
					ENSMUSG00000036173	HISTONE H3 (H3.2).
132	13.23170125-23173292	77.3	0.0622	0.3591	ENSMUSG00000036132	HISTONE H1.1 (H1 VAR.3) (H1A).
					ENSMUSG00000036173	HISTONE H3 (H3.2).
133	13.27591935-27593045	41.0	0.0119	0.3837	ENSMUSG00000017064	NA
134	13.52751878-52756081	289.4	0.0930	2.6253	ENSMUSG00000021469	HOMEOBOX PROTEIN MSX-2 (HOX-8.1).
135	13.55311592-55317979	254.6	0.0955	2.1841	ENSMUSG00000021506	PITUITARY HOMEOBOX 1 (HOMEOBOX PROTEIN P-OTX) (PITUITARY OTX-RELATED FACTOR) (HINDLIMB EXPRESSED HOMEOBOX PROTEIN BACKFOOT) (PTX1).
136	13.60496444-60496700	32.0	0.0123	0.2722	ENSMUSG00000021433	CTLA-2-BETA PROTEIN PRECURSOR (FRAGMENT).
					ENSMUSG00000033834	TROPHOBLAST-SPECIFIC PROTEIN PRECURSOR.
137	13.62386305-62396461	335.9	0.0795	3.3092	ENSMUSG00000021466	PATCHED PROTEIN HOMOLOG 1 (PTC1) (PTC).
138	13.69029605-69034427	94.9	0.0219	0.9398	ENSMUSG00000021602	IROQUOIS-CLASS HOMEODOMAIN PROTEIN IRX-1 (FRAGMENT).
139	13.70335945-70338999	91.1	0.0359	0.7664	ENSMUSG00000021604	IROQUOIS RELATED HOMEOBOX 4 (DROSOPHILA).
140	13.75280509-75289493	125.8	0.0364	1.1782	ENSMUSG00000035892	COUP TRANSCRIPTION FACTOR 1 (COUP-TF1) (COUP-TF I).
141	13.75295999-75296808	24.2	0.0208	0.0995	ENSMUSG00000035892	COUP TRANSCRIPTION FACTOR 1 (COUP-TF1) (COUP-TF I).
142	13.88157413-88157619	32.0	0.0397	0.0199	ENSMUSG00000021619	GENE.
143	13.92203897-92209278	320.1	0.0860	3.0592	ENSMUSG00000021685	HOMEOBOX PROTEIN ORTHOPEDIA.
144	13.97261000-97264806	334.5	0.0858	3.2335	ENSMUSG00000021647	COCAINE- AND AMPHETAMINE-REGULATED TRANSCRIPT PROTEIN PRECURSOR
145	14.108328953-108329184	33.1	0.0361	0.0665	ENSMUSG00000032891	GLYPICAN-6 PRECURSOR.
146	14.108529702-108529910	24.0	0.0217	0.0894	ENSMUSG00000032891	GLYPICAN-6 PRECURSOR.
147	14.110165101-110167920	88.0	0.0852	0.2761	ENSMUSG00000003953	NA
					ENSMUSG00000022133	NA
					ENSMUSG00000022136	DNAJ (HSP40) HOMOLOG, SUBFAMILY C, MEMBER

						3
148	14.110692385-110692613	19.5	0.0233	0.0201	ENSMUSG00000042021	HEPARAN SULFATE 6-O-SULFOTRANSFERASE 3.
149	14.113365106-113368030	137.8	0.1231	0.5262	ENSMUSG00000025544	TRANSMEMBRANE 9 SUPERFAMILY PROTEIN MEMBER 2 PRECURSOR.
150	14.115910243-115910519	29.5	0.0184	0.1856	ENSMUSG00000025551	FIBROBLAST GROWTH FACTOR-14 (FGF-14) (FIBROBLAST GROWTH FACTOR HOMOLOGOUS FACTOR 4) (FHF-4).
151	14.17101881-17113285	430.8	0.1062	4.2038	ENSMUSG00000021778	NA
152	14.23611440-23618208	147.0	0.0444	1.3596	ENSMUSG00000021994	WNT-5A PROTEIN PRECURSOR.
153	14.44583671-44584386	17.2	0.0033	0.1771	ENSMUSG00000022164	T-CELL RECEPTOR ALPHA CHAIN V REGION 2B4 PRECURSOR.
154	14.49205978-49210930	77.4	0.0590	0.3889	ENSMUSG00000021974	GLIA-ACTIVATING FACTOR PRECURSOR (GAF) (FIBROBLAST GROWTH FACTOR-9) (FGF-9) (HBGF-9).
155	14.58366810-58371481	135.6	0.1121	0.6024	ENSMUSG00000022053	TRANSCRIPTION FACTOR COE2 (EARLY B-CELL FACTOR 2) (EBF-2) (OLF-1/EBF- LIKE 3) (OE-3) (O/E-3) (METENCEPHALON-MESENCEPHALON-OLFACTORY TRANSCRIPTION FACTOR 1) (MET-MESENCEPHALON-OLFACTORY TRANSCRIPTION FACTOR 1) (MET-MESENCEPHAL
156	14.70755874-70761803	88.8	0.0998	0.1523	ENSMUSG00000036422	PROTOCADHERIN 8.
157	14.81042676-81043614	30.7	0.0211	0.1758	ENSMUSG00000035043	SIMILAR TO POLY(A)-BINDING PROTEIN, CYTOPLASMIC 4 (INDUCIBLE FORM).
158	14.9636744-9637251	19.1	0.0100	0.1388	ENSMUSG00000021734	INTERLEUKIN-3 RECEPTOR CLASS II ALPHA CHAIN PRECURSOR.
159	14.9637808-9638354	11.9	0.0100	0.0512	ENSMUSG00000021734	INTERLEUKIN-3 RECEPTOR CLASS II ALPHA CHAIN PRECURSOR.
160	14.9643298-9643587	9.7	0.0100	0.0250	ENSMUSG00000021734	INTERLEUKIN-3 RECEPTOR CLASS II ALPHA CHAIN PRECURSOR.
161	15.103911997-103914453	55.3	0.0505	0.2014	ENSMUSG00000001656	HOMEOBOX PROTEIN HOX-C12 (HOX-3.8) (FRAGMENT).
					ENSMUSG00000022484	HOMEOBOX PROTEIN HOX-C10 (HOX-3.6).
					ENSMUSG00000036139	HOMEOBOX PROTEIN HOX-C9 (HOX-3.2).
162	15.103931300-103932309	8.5	0.0100	0.0097	ENSMUSG00000001656	HOMEOBOX PROTEIN HOX-C12 (HOX-3.8) (FRAGMENT).
					ENSMUSG00000001657	HOMEOBOX PROTEIN HOX-C8 (HOX-3.1) (M31).
					ENSMUSG00000022484	HOMEOBOX PROTEIN HOX-C10 (HOX-3.6).
					ENSMUSG00000036139	HOMEOBOX PROTEIN HOX-C9 (HOX-3.2).
163	15.103943516-103943743	12.4	0.0008	0.1416	ENSMUSG00000001657	HOMEOBOX PROTEIN HOX-C8 (HOX-3.1) (M31).
					ENSMUSG00000001661	HOMEOBOX PROTEIN HOX-C6 (HOX-3.3) (HOX-6.1).
					ENSMUSG00000022484	HOMEOBOX PROTEIN HOX-C10 (HOX-3.6).
					ENSMUSG00000022485	HOMEOBOX PROTEIN HOX-C5 (HOX-3.4) (HOX-6.2).
					ENSMUSG00000036139	HOMEOBOX PROTEIN HOX-C9 (HOX-3.2).
164	15.103949349-103949573	1.7	0.0008	0.0133	ENSMUSG00000001657	HOMEOBOX PROTEIN HOX-C8 (HOX-3.1) (M31).
					ENSMUSG00000001661	HOMEOBOX PROTEIN HOX-C6 (HOX-3.3) (HOX-6.1).
					ENSMUSG00000022485	HOMEOBOX PROTEIN HOX-C5 (HOX-3.4) (HOX-6.2).
					ENSMUSG00000036139	HOMEOBOX PROTEIN HOX-C9 (HOX-3.2).
165	15.103953855-103954086	6.7	0.0011	0.0705	ENSMUSG00000001657	HOMEOBOX PROTEIN HOX-C8 (HOX-3.1) (M31).
					ENSMUSG00000001661	HOMEOBOX PROTEIN HOX-C6 (HOX-3.3) (HOX-6.1).
					ENSMUSG00000022485	HOMEOBOX PROTEIN HOX-C5 (HOX-3.4) (HOX-6.2).

					ENSMUSG00000036139	HOMEOBOX PROTEIN HOX-C9 (HOX-3.2).
166	15.103986584-103987987	74.8	0.0214	0.7034	ENSMUSG00000001661	HOMEOBOX PROTEIN HOX-C6 (HOX-3.3) (HOX-6.1).
					ENSMUSG00000022485	HOMEOBOX PROTEIN HOX-C5 (HOX-3.4) (HOX-6.2).
					ENSMUSG00000022486	HOMEOBOX PROTEIN HOX-C4 (HOX-3.5).
167	15.10598728-10600168	83.1	0.0762	0.3001	ENSMUSG00000022246	ANKYCORBIN
					ENSMUSG00000022249	NA
168	15.19053083-19053367	29.6	0.0296	0.0843	ENSMUSG00000022321	CADHERIN-10 (T2-CADHERIN) (FRAGMENT).
169	15.35393215-35397100	127.4	0.0794	0.8032	ENSMUSG00000022330	ODD-SKIPPED-RELATED 2 PROTEIN.
170	15.7632763-7638457	62.8	0.0477	0.3174	ENSMUSG00000022144	GLIAL CELL LINE-DERIVED NEUROTROPHIC FACTOR PRECURSOR.
171	15.97339477-97343786	98.9	0.1043	0.2318	ENSMUSG00000033228	NA
172	15.98911550-98914110	94.3	0.0755	0.4413	ENSMUSG00000022483	COLLAGEN ALPHA 1(II) CHAIN PRECURSOR
173	16.19037149-19037370	11.1	0.0124	0.0202	ENSMUSG00000022695	IG LAMBDA-2 CHAIN V REGION PRECURSOR.
174	16.23744310-23746176	67.3	0.0691	0.1751	ENSMUSG00000022508	B-CELL LYMPHOMA 6 PROTEIN HOMOLOG.
175	16.58693209-58693471	95.9	0.0548	0.6501	ENSMUSG00000035107	NA
176	16.76309614-76310261	30.1	0.0246	0.1359	ENSMUSG00000032932	NA
177	16.89421934-89422163	16.8	0.0205	0.0140	ENSMUSG00000040970	NA
					ENSMUSG00000040984	KERATIN-ASSOCIATED PROTEIN 13.
178	16.89431426-89431787	29.5	0.0175	0.1945	ENSMUSG00000040970	NA
					ENSMUSG00000040984	KERATIN-ASSOCIATED PROTEIN 13.
179	16.10022477-10025712	130.3	0.1310	0.3646	ENSMUSG00000038055	MYLE PROTEIN (DEXAMETHASONE-INDUCED PROTEIN).
180	17.13516524-13516735	16.3	0.0202	0.0108	ENSMUSG00000023886	SECRETED MODULAR CALCIUM-BINDING PROTEIN 2.
181	17.31065239-31070426	260.3	0.0352	2.8059	ENSMUSG00000024042	PROBABLE SERINE/THREONINE PROTEIN KINASE SNF1LK (EC 2.7.1.-) (HRT-20) (MYOCARDIAL SNF1-LIKE KINASE).
182	17.31249825-31252102	107.2	0.1327	0.0712	ENSMUSG00000002070	NA
					ENSMUSG00000002076	NA
183	17.35193993-35194721	37.6	0.0464	0.0263	ENSMUSG00000024432	H-2 CLASS I HISTOCOMPATIBILITY ANTIGEN, TLA(B) ALPHA CHAIN PRECURSOR (MHC THYMUS LEUKEMIA ANTIGEN).
					ENSMUSG00000038311	NA
184	17.54741202-54741531	21.3	0.0221	0.0538	ENSMUSG00000024174	NA
185	17.55481990-55485168	52.9	0.0531	0.1487	ENSMUSG00000013236	PROTEIN-TYROSINE PHOSPHATASE, RECEPTOR-TYPE, S PRECURSOR (EC 3.1.3.48) (PROTEIN-TYROSINE PHOSPHATASE SIGMA) (RPTP-SIGMA) (PROTEIN TYROSINE PHOSPHATASE PTPT9) (PTPASE NU-3).
					ENSMUSG00000024201	SIMILAR TO KIAA0677 GENE PRODUCT.
					ENSMUSG00000024203	NA
186	18.37117509-37119362	91.7	0.0939	0.2407	ENSMUSG00000007705	PROTOCADHERIN ALPHA 4
					ENSMUSG00000007706	CADHERIN-RELATED NEURAL RECEPTER 6 (FRAGMENT).
					ENSMUSG00000007707	CADHERIN-RELATED NEURAL RECEPTER 4 (FRAGMENT).

					ENSMUSG00000041677	PROTOCADHERIN ALPHA 6
					ENSMUSG00000041804	PROTOCADHERIN ALPHA 3.
					ENSMUSG00000041843	PROTOCADHERIN ALPHA 2.
					ENSMUSG00000041884	PROTOCADHERIN ALPHA 1.
187	18.37126631-37128370	130.7	0.0935	0.7131	ENSMUSG00000007705	PROTOCADHERIN ALPHA 4
					ENSMUSG00000007706	CADHERIN-RELATED NEURAL RECEPTER 6 (FRAGMENT).
					ENSMUSG00000007707	CADHERIN-RELATED NEURAL RECEPTER 4 (FRAGMENT).
					ENSMUSG00000041677	PROTOCADHERIN ALPHA 6
					ENSMUSG00000041804	PROTOCADHERIN ALPHA 3.
					ENSMUSG00000041843	PROTOCADHERIN ALPHA 2.
188	18.37139991-37141467	83.8	0.1011	0.0799	ENSMUSG00000007705	PROTOCADHERIN ALPHA 4
					ENSMUSG00000007706	CADHERIN-RELATED NEURAL RECEPTER 6 (FRAGMENT).
					ENSMUSG00000007707	CADHERIN-RELATED NEURAL RECEPTER 4 (FRAGMENT).
					ENSMUSG00000041564	CADHERIN-RELATED NEURAL RECEPTER 8 (FRAGMENT).
					ENSMUSG00000041599	PROTOCADHERIN ALPHA 8.
					ENSMUSG00000041677	PROTOCADHERIN ALPHA 6
189	18.37161117-37163029	95.5	0.0948	0.2790	ENSMUSG00000007440	PROTOCADHERIN ALPHA C2.
					ENSMUSG00000007707	CADHERIN-RELATED NEURAL RECEPTER 4 (FRAGMENT).
					ENSMUSG00000041467	CADHERIN-RELATED NEURAL RECEPTER 7 (FRAGMENT).
					ENSMUSG00000041508	CADHERIN-RELATED NEURAL RECEPTER 3 (FRAGMENT).
					ENSMUSG00000041564	CADHERIN-RELATED NEURAL RECEPTER 8 (FRAGMENT).
					ENSMUSG00000041599	PROTOCADHERIN ALPHA 8.
190	18.37192627-37194006	70.0	0.0644	0.2501	ENSMUSG00000007440	PROTOCADHERIN ALPHA C2.
					ENSMUSG00000041467	CADHERIN-RELATED NEURAL RECEPTER 7 (FRAGMENT).
					ENSMUSG00000041508	CADHERIN-RELATED NEURAL RECEPTER 3 (FRAGMENT).
191	18.37207450-37209321	88.5	0.0418	0.6805	ENSMUSG00000007440	PROTOCADHERIN ALPHA C2.
192	18.37917848-37920059	129.7	0.1062	0.5846	ENSMUSG00000024463	PROTOCADHERIN 13.
					ENSMUSG00000039718	PROTOCADHERIN GAMMA A11.
					ENSMUSG00000039802	PROTOCADHERIN GAMMA B7.
					ENSMUSG00000039839	PROTOCADHERIN GAMMA A10.
					ENSMUSG00000039882	PROTOCADHERIN GAMMA A9.
					ENSMUSG00000039933	PROTOCADHERIN GAMMA B5.
193	18.37929138-37930971	126.2	0.1279	0.3433	ENSMUSG00000024463	PROTOCADHERIN 13.
					ENSMUSG00000039718	PROTOCADHERIN GAMMA A11.
					ENSMUSG00000039802	PROTOCADHERIN GAMMA B7.
					ENSMUSG00000039839	PROTOCADHERIN GAMMA A10.
194	18.63501432-63502121	40.0	0.0378	0.1345	ENSMUSG00000040908	NA

195	18.7178424-7179004	41.2	0.0334	0.1889	ENSMUSG00000024280	NA
196	19.12427634-12428001	26.9	0.0340	0.0112	ENSMUSG00000038666	NA
197	19.39769328-39769573	18.2	0.0081	0.1450	ENSMUSG00000042248	SIMILAR TO CYTOCHROME P450, 2C37.
198	19.43499593-43504124	126.9	0.1334	0.3019	ENSMUSG00000025191	HOMEOBOX PROTEIN NKX-2.3.
199	19.56855416-56860208	299.2	0.0691	2.9629	ENSMUSG00000025081	TUDOR DOMAIN CONTAINING PROTEIN 1.
200	19.61327560-61328116	22.4	0.0279	0.0133	ENSMUSG00000041980	GRANULOCYTE-MACROPHAGE COLONY-STIMULATING FACTOR RECEPTOR ALPHA CHAIN PRECURSOR (GM-CSF-R-ALPHA) (GMR).
201	Un.479515-479818	24.0	0.0091	0.2055	ENSMUSG00000004024	EUKARYOTIC TRANSLATION INITIATION FACTOR 1A (EIF-1A) (EIF-4C) (FRAGMENT).
202	Un.87409427-87409678	17.2	0.0177	0.0444	ENSMUSG00000021888	VOMERONASAL 2, RECEPTOR, 11
203	Un.48665809-48666040	16.4	0.0209	0.0054	ENSMUSG00000024674	NA
204	Un.39514237-39514762	30.1	0.0227	0.1535	ENSMUSG00000033412	T-CELL RECEPTOR ALPHA CHAIN V REGION PHDS58 PRECURSOR.
205	Un.35079087-35079499	28.9	0.0199	0.1641	ENSMUSG00000033431	NA
206	Un.35112573-35112843	20.4	0.0203	0.0591	ENSMUSG00000033431	NA
207	Un.35939364-35939637	33.2	0.0337	0.0900	ENSMUSG00000033431	NA
208	Un.36106885-36107112	21.6	0.0208	0.0693	ENSMUSG00000033431	NA
209	Un.36107737-36108176	19.3	0.0207	0.0415	ENSMUSG00000033431	NA
210	Un.31903807-31904298	25.7	0.0240	0.0892	ENSMUSG00000033563	FERRITIN LIGHT CHAIN 2 (FERRITIN L SUBUNIT 2) (FERRITIN SUBUNIT LG).
211	Un.29259500-29259785	11.4	0.0106	0.0399	ENSMUSG00000033647	60S RIBOSOMAL PROTEIN L30.
212	Un.21587787-21588546	18.8	0.0209	0.0342	ENSMUSG00000033844	UNKNOWN (PROTEIN FOR MGC=6827).
213	Un.107047921-107048175	20.3	0.0237	0.0261	ENSMUSG00000035261	NA
214	Un.106292066-106297686	230.9	0.0356	2.4487	ENSMUSG00000035312	ADULT MALE STOMACH CDNA, RIKEN FULL-LENGTH ENRICHED LIBRARY, CLONE=2210018M05, FULL INSERT SEQUENCE (FRAGMENT).
215	Un.101510780-101511277	28.7	0.0345	0.0290	ENSMUSG00000035482	RING1 AND YY1 BINDING PROTEIN
216	Un.19820425-19821185	18.8	0.0177	0.0636	ENSMUSG00000035684	UNKNOWN (PROTEIN FOR MGC=6827).
217	Un.14128600-14128805	21.5	0.0184	0.0897	ENSMUSG00000035746	NA
218	Un.45133993-45134246	15.1	0.0135	0.0578	ENSMUSG00000042132	HISTONE H2B F (H2B 291A).

APPENDICES

Appendix 1: Frequencies of Significant Heptamers at the level $\alpha = 10^{-5}$

Heptamer	DMR-CGIs	Control CGIs	LOD	*p*-value
ACACACA	0.0002339	0.0006223	-1.41	2.80E-06
AGGCGGC	0.0000439	0.0003064	-2.80	1.39E-06
CACACAC	0.0003070	0.0007220	-1.23	4.30E-06
CCGCCGC	0.0002485	0.0009788	-1.98	3.28E-13
CGCCGCC	0.0002485	0.0009383	-1.92	2.60E-12
CGCCGCG	0.0001316	0.0004528	-1.78	3.52E-06
CGCGGCG	0.0001316	0.0004528	-1.78	3.52E-06
CGGCGGC	0.0001900	0.0008085	-2.09	8.78E-12
GCCGCCG	0.0001900	0.0008085	-2.09	8.78E-12
GCCGCCT	0.0000439	0.0003064	-2.80	1.39E-06
GCGCCGC	0.0001754	0.0005408	-1.62	1.69E-06
GCGGCGC	0.0001754	0.0005408	-1.62	1.69E-06
GCGGCGG	0.0002485	0.0009788	-1.98	3.28E-13
GGCGGCG	0.0002485	0.0009383	-1.92	2.60E-12
GTGTGTG	0.0003070	0.0007220	-1.23	4.30E-06
TGTGTGT	0.0002339	0.0006223	-1.41	2.80E-06
AAAATTT	0.0001462	0.0000230	2.67	8.81E-07
AAATTTG	0.0001608	0.0000168	3.26	3.73E-09
AAATTTT	0.0001462	0.0000230	2.67	8.81E-07
AAGCCCT	0.0002631	0.0000728	1.85	1.33E-06
ACAACTC	0.0001754	0.0000408	2.10	7.64E-06
ACAATGC	0.0001462	0.0000293	2.32	8.45E-06
ACTAATG	0.0001169	0.0000091	3.68	2.25E-08
AGCAGGG	0.0003801	0.0001454	1.39	5.81E-06
AGCCGAG	0.0008332	0.0001789	2.22	3.12E-21
AGGGCTT	0.0002631	0.0000728	1.85	1.33E-06
ATGAGCA	0.0001900	0.0000381	2.32	6.86E-07
ATGTGAC	0.0002339	0.0000246	3.25	3.90E-12
ATTAGTA	0.0001023	0.0000064	4.00	2.22E-08
ATTGCAA	0.0001169	0.0000172	2.77	4.06E-06
ATTTAGG	0.0001316	0.0000216	2.61	3.59E-06
CAAATTT	0.0001608	0.0000168	3.26	3.73E-09
CAACACG	0.0001608	0.0000327	2.30	4.23E-06
CAGCCGA	0.0003070	0.0000959	1.68	1.64E-06
CATGTGA	0.0001900	0.0000249	2.93	4.08E-09
CATTAGT	0.0001169	0.0000091	3.68	2.25E-08
CCCCCAA	0.0003655	0.0001157	1.66	2.90E-07
CCCCCCC	0.0012133	0.0003368	1.85	1.07E-22
CCCTGCT	0.0003801	0.0001454	1.39	5.81E-06

CCGAGCC	0.0008186	0.0002751	1.57	1.08E-12
CCTAAAT	0.0001316	0.0000216	2.61	3.59E-06
CCTGCTC	0.0004970	0.0001958	1.34	6.41E-07
CGAATGA	0.0001169	0.0000173	2.75	4.40E-06
CGAGCCG	0.0004970	0.0001869	1.41	2.26E-07
CGCCAAC	0.0002339	0.0000656	1.83	5.07E-06
CGCGGCA	0.0004532	0.0001309	1.79	1.95E-09
CGGCTCG	0.0004970	0.0001869	1.41	2.26E-07
CGTGTTG	0.0001608	0.0000327	2.30	4.23E-06
CTCGGCT	0.0008332	0.0001789	2.22	3.12E-21
GAGCAGG	0.0004970	0.0001958	1.34	6.41E-07
GAGCCGA	0.0005409	0.0001335	2.02	8.51E-13
GAGTTGT	0.0001754	0.0000408	2.10	7.64E-06
GCATTGT	0.0001462	0.0000293	2.32	8.45E-06
GCCGAGC	0.0009648	0.0002779	1.80	1.00E-17
GCGCCAA	0.0002777	0.0000818	1.76	1.86E-06
GCGGCAA	0.0003508	0.0000871	2.01	5.16E-09
GCTCGGC	0.0009648	0.0002779	1.80	1.00E-17
GGCTCGG	0.0008186	0.0002751	1.57	1.08E-12
GGGGGGG	0.0012133	0.0003368	1.85	1.07E-22
GTCACAT	0.0002339	0.0000246	3.25	3.90E-12
GTGACAA	0.0002777	0.0000349	2.99	1.56E-12
GTTGGCG	0.0002339	0.0000656	1.83	5.07E-06
GTTGTCA	0.0002339	0.0000317	2.88	1.93E-10
TACTAAT	0.0001023	0.0000064	4.00	2.22E-08
TAGTACA	0.0000877	0.0000100	3.13	7.81E-06
TCACATG	0.0001900	0.0000249	2.93	4.08E-09
TCATTCG	0.0001169	0.0000173	2.75	4.40E-06
TCGGCTC	0.0005409	0.0001335	2.02	8.51E-13
TCGGCTG	0.0003070	0.0000959	1.68	1.64E-06
TGACAAC	0.0002339	0.0000317	2.88	1.93E-10
TGCCGCA	0.0003216	0.0000868	1.89	8.35E-08
TGCCGCG	0.0004532	0.0001309	1.79	1.95E-09
TGCGGCA	0.0003216	0.0000868	1.89	8.35E-08
TGCTCAT	0.0001900	0.0000381	2.32	6.86E-07
TGTACTA	0.0000877	0.0000100	3.13	7.81E-06
TGTCACA	0.0002193	0.0000332	2.72	2.80E-09
TGTGACA	0.0002193	0.0000332	2.72	2.80E-09
TTGCAAT	0.0001169	0.0000172	2.77	4.06E-06
TTGCCGC	0.0003508	0.0000871	2.01	5.16E-09
TTGGCGC	0.0002777	0.0000818	1.76	1.86E-06
TTGGGGG	0.0003655	0.0001157	1.66	2.90E-07
TTGTCAC	0.0002777	0.0000349	2.99	1.56E-12

Appendix 2: Predicted Novel DMR-CGIs

The genomic coordinates, CGI Score, regional SINE Content, LDA value, EnsEMBL identifier, and EnsEMBL gene description for all of the 218 predicted novel DMR-CGIs are represented. EnsEMBL gene loci that are associated with FANTOM2 Candidate transcripts are highlighted in green.

#	CGI Locus	Score	SINE	LDA	EnsEMBL Gene ID	Gene Description
1	1.107929988-107930208	13.1	0.0167	0.0043	ENSMUSG00000026327	NA
					ENSMUSG00000026328	18 DAYS EMBRYO CDNA, RIKEN FULL-LENGTH ENRICHED LIBRARY, CLONE=1110013A16, FULL INSERT SEQUENCE.
2	1.139715233-139715939	55.5	0.0474	0.2313	ENSMUSG00000019230	LIM/HOMEOBOX PROTEIN LHX9.
3	1.140669281-140669512	24.1	0.0082	0.2140	ENSMUSG00000026362	COMPLEMENT FACTOR H-RELATED PROTEIN.
4	1.140750137-140750383	8.1	0.0089	0.0165	ENSMUSG00000033898	COMPLEMENT FACTOR H-RELATED PROTEIN.
5	1.145642591-145642940	23.1	0.0260	0.0390	ENSMUSG00000026357	REGULATOR OF G-PROTEIN SIGNALING 18 (RGS18).
6	1.167571950-167572874	91.7	0.0719	0.4435	ENSMUSG00000026563	HYPOTHETICAL 37.4 KDA PROTEIN.
7	1.168898960-168903083	180.0	0.0636	1.5803	ENSMUSG00000026686	LIM HOMEODOMAIN-CONTAINING TRANSCRIPTION FACTOR.
8	1.191520093-191526544	113.4	0.0617	0.7975	ENSMUSG00000010175	HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1).
9	1.34631926-34633988	59.0	0.0641	0.1208	ENSMUSG00000026131	BULLOUS PEMPHIGOID ANTIGEN 1-B.
10	1.44310462-44310729	40.7	0.0444	0.0818	ENSMUSG00000026052	NA
11	1.44313593-44313893	40.4	0.0417	0.1036	ENSMUSG00000026052	NA
12	1.4500872-4506849	255.4	0.0745	2.3870	ENSMUSG00000025902	TRANSCRIPTION FACTOR SOX-17.
13	1.5575123-5575413	20.2	0.0196	0.0629	ENSMUSG00000025905	KAPPA-TYPE OPIOID RECEPTOR (KOR-1) (MSL-1).
14	1.76346798-76349491	92.9	0.0979	0.2184	ENSMUSG00000006576	ANION EXCHANGE PROTEIN 3 (NEURONAL BAND 3-LIKE PROTEIN).
					ENSMUSG00000026213	NA
					ENSMUSG00000032968	INHIBIN ALPHA CHAIN PRECURSOR.
15	2.110630098-110632418	49.9	0.0464	0.1734	ENSMUSG00000041693	BRAIN-DERIVED NEUROTROPHIC FACTOR PRECURSOR (BDNF).
16	2.110646455-110648177	37.6	0.0487	0.0054	ENSMUSG00000041693	BRAIN-DERIVED NEUROTROPHIC FACTOR PRECURSOR (BDNF).
17	2.148392152-148395933	143.6	0.0638	1.1409	ENSMUSG00000037034	PAIRED BOX PROTEIN PAX-1.
18	2.149075615-149080794	175.8	0.0661	1.5063	ENSMUSG00000037025	HEPATOCYTE NUCLEAR FACTOR 3-BETA (HNF-3B).
19	2.149441396-149442253	48.1	0.0582	0.0439	ENSMUSG00000027437	THROMBOMODULIN PRECURSOR (FETOMODULIN) (TM).
20	2.162898605-162901835	158.3	0.0652	1.3042	ENSMUSG00000027413	PROTEIN TYROSINE PHOSPHATASE, RECEPTOR TYPE, T.
21	2.165117351-165120836	78.8	0.0877	0.1429	ENSMUSG00000040164	K+ VOLTAGE-GATED CHANNEL, SUBFAMILY S, 1.
22	2.18817259-18822395	189.8	0.0831	1.5189	ENSMUSG00000026739	POLYCOMB COMPLEX PROTEIN BMI-1.
23	2.25854939-25860390	121.7	0.1335	0.2388	ENSMUSG00000026942	TNF RECEPTOR ASSOCIATED FACTOR 2 (TRAF2).

					ENSMUSG00000036504	NA
24	2.55961454-55962223	42.2	0.0382	0.1562	ENSMUSG00000026824	G PROTEIN-ACTIVATED INWARD RECTIFIER POTASSIUM CHANNEL 1 (GIRK1) (POTASSIUM CHANNEL, INWARDLY RECTIFYING, SUBFAMILY J, MEMBER 3) (INWARD RECTIFIER K+ CHANNEL KIR3.1) (KGA) (KGB1).
25	2.57640962-57642961	101.9	0.0699	0.5836	ENSMUSG00000026826	ORPHAN NUCLEAR RECEPTOR NURR1 (NUR-RELATED FACTOR 1).
26	2.72320504-72325498	130.5	0.1218	0.4514	ENSMUSG00000041911	HOMEOBOX PROTEIN DLX-1.
27	2.74106634-74116088	442.8	0.1133	4.2829	ENSMUSG00000008226	NA
28	2.75516613-75530035	613.6	0.0701	6.7331	ENSMUSG00000001817	HOMEOBOX PROTEIN HOX-D10 (HOX-4.5) (HOX-5.3).
					ENSMUSG00000001819	HOMEOBOX PROTEIN HOX-D13 (HOX-4.8).
					ENSMUSG00000001823	HOMEOBOX PROTEIN HOX-D12 (HOX-4.7) (HOX-5.6).
					ENSMUSG00000042482	HOMEOBOX PROTEIN HOX-D9 (HOX-4.4) (HOX-5.2).
					ENSMUSG00000042499	HOMEOBOX PROTEIN HOX-D11 (HOX-4.6) (HOX-5.5).
29	2.75543059-75555527	644.5	0.0365	7.4138	ENSMUSG00000001815	HOMEOBOX EVEN-SKIPPED HOMOLOG PROTEIN 2 (EVX-2).
					ENSMUSG00000027102	HOMEOBOX PROTEIN HOX-D8 (HOX-4.3) (HOX-5.4).
					ENSMUSG00000042464	HOMEOBOX PROTEIN HOX-D4 (HOX-4.2) (HOX-5.1).
					ENSMUSG00000042482	HOMEOBOX PROTEIN HOX-D9 (HOX-4.4) (HOX-5.2).
30	2.75565849-75573593	83.9	0.0230	0.7972	ENSMUSG00000042453	HOMEOBOX PROTEIN HOX-D3 (HOX-4.1) (MH-19).
31	2.80334667-80337688	194.7	0.0355	2.0146	ENSMUSG00000034701	NEUROGENIC DIFFERENTIATION FACTOR 1 (NEUROD1).
32	2.84864734-84865841	47.4	0.0394	0.2088	ENSMUSG00000034552	NA
33	2.94622704-94627599	110.6	0.0672	0.7132	ENSMUSG00000040310	HOMEOBOX PROTEIN ARISTALESS-LIKE 4 (ALX-4).
34	2.9912809-9918564	211.8	0.1344	1.3131	ENSMUSG00000015619	TRANS-ACTING T-CELL SPECIFIC TRANSCRIPTION FACTOR GATA-3.
					ENSMUSG00000025783	NA
35	3.129708234-129713310	173.3	0.0610	1.5237	ENSMUSG00000028023	PITUITARY HOMEOBOX 2 (ORTHODENTICLE-LIKE HOMEOBOX 2) (SOLURSHIN) (ALL1 RESPONSIVE PROTEIN ARP1) (BRX1 HOMEOPROTEIN) (PAIRED-LIKE HOMEODOMAIN TRANSCRIPTION FACTOR MUNC 30).
36	3.131661313-131667880	96.6	0.1087	0.1640	ENSMUSG00000027985	LYMPHOID ENHANCER BINDING FACTOR 1 (LEF-1).
37	3.154910223-154915320	308.0	0.0688	3.0720	ENSMUSG00000028201	LIM/HOMEOBOX PROTEIN LHX8 (L3).
38	3.17768553-17771839	35.8	0.0335	0.1223	ENSMUSG00000039527	BASIC HELIX-LOOP-HELIX DOMAIN CONTAINING, CLASS B5
39	3.29825267-29826282	53.1	0.0388	0.2819	ENSMUSG00000027684	ECOTROPIC VIRUS INTEGRATION 1 SITE PROTEIN.
40	3.45477661-45482475	184.8	0.0355	1.8952	ENSMUSG00000027730	NA
					ENSMUSG00000037927	PROTOCADHERIN 10
41	3.55275165-55278987	148.6	0.0974	0.8924	ENSMUSG00000036615	REGULATORY FACTOR X-ASSOCIATED PROTEIN.
42	3.60382842-60383258	24.1	0.0056	0.2384	ENSMUSG00000036976	NA
43	3.67580670-67584119	79.9	0.0740	0.2814	ENSMUSG00000027833	SHORT STATURE HOMEOBOX PROTEIN 2 (HOMEOBOX PROTEIN OG12X) (OG-12) (PAIRED FAMILY HOMEODOMAIN PROTEIN PRX3).
					ENSMUSG00000034544	NA

44	3.96544118-96552984	384.2	0.1066	3.6401	ENSMUSG00000038445	HISTONE H3 (H3.2).
45	3.96791854-96794377	100.1	0.1132	0.1640	ENSMUSG00000038403	NA
46	4.11571701-11574111	147.0	0.0577	1.2375	ENSMUSG00000040642	NA
47	4.43550588-43552290	117.2	0.1264	0.2500	ENSMUSG00000014030	PAIRED BOX PROTEIN PAX-5 (B-CELL SPECIFIC TRANSCRIPTION FACTOR) (BSAP).
48	4.51921727-51922560	65.4	0.0438	0.3836	ENSMUSG00000015243	ATP-BINDING CASSETTE, SUB-FAMILY A, MEMBER 1 (ATP-BINDING CASSETTE TRANSPORTER 1) (ATP-BINDING CASSETTE 1) (ABC-1).
49	4.79784817-79794020	243.5	0.0544	2.4289	ENSMUSG00000008575	NUCLEAR FACTOR 1 B-TYPE (NUCLEAR FACTOR 1/B) (NF1-B) (NFI-B) (NF-I/B) (CCAAT-BOX BINDING TRANSCRIPTION FACTOR) (CTF) (TGGCA-BINDING PROTEIN).
50	4.93781858-93782066	22.1	0.0268	0.0190	ENSMUSG00000028571	SIMILAR TO CYP2J4.
51	5.105394792-105400633	143.0	0.1236	0.5851	ENSMUSG00000029275	ZINC FINGER PROTEIN GFI-1 (GROWTH FACTOR INDEPENDENCE-1).
52	5.127399336-127400484	88.3	0.0913	0.2243	ENSMUSG00000034268	40S RIBOSOMAL PROTEIN S16.
53	5.18627019-18627391	28.6	0.0246	0.1184	ENSMUSG00000040003	ACTIVIN RECEPTOR INTERACTING PROTEIN 1.
54	5.20644740-20645372	30.7	0.0372	0.0279	ENSMUSG00000038490	REELIN PRECURSOR (EC 3.4.21.-) (REELER PROTEIN).
55	5.34282539-34284302	115.4	0.0576	0.8581	ENSMUSG00000029097	NA
					ENSMUSG00000029098	ACYL-COENZYME A OXIDASE 3, PEROXISOMAL (EC 1.3.3.6) (PRISTANOYL-COA OXIDASE).
56	5.36724455-36728222	107.3	0.0783	0.5719	ENSMUSG00000039404	HOMEOBOX PROTEIN MSX-1 (HOX-7) (HOX-7.1).
57	5.40734938-40740238	268.3	0.0420	2.8401	ENSMUSG00000029129	HOMEOBOX PROTEIN NKX-3.2 (BAGPIPE HOMEOBOX PROTEIN HOMOLOG 1).
58	5.70959108-70959357	21.8	0.0228	0.0532	ENSMUSG00000029212	GAMMA-AMINOBUTYRIC-ACID RECEPTOR BETA-1 SUBUNIT PRECURSOR (GABA(A) RECEPTOR).
59	5.71184005-71184388	27.9	0.0269	0.0879	ENSMUSG00000029212	GAMMA-AMINOBUTYRIC-ACID RECEPTOR BETA-1 SUBUNIT PRECURSOR (GABA(A) RECEPTOR).
60	5.95527555-95528188	36.0	0.0378	0.0860	ENSMUSG00000029338	HYPOTHETICAL 24.5 KDA PROTEIN (FRAGMENT).
61	6.127436907-127437245	39.7	0.0465	0.0512	ENSMUSG00000038097	VOLTAGE-GATED POTASSIUM CHANNEL PROTEIN KV1.5 (KV1-5).
62	6.127540710-127542681	51.5	0.0620	0.0502	ENSMUSG00000003015	VOLTAGE-GATED POTASSIUM CHANNEL PROTEIN KV1.1 (MK1) (MBK1).
63	6.129836345-129836967	29.2	0.0291	0.0843	ENSMUSG00000000248	NA
64	6.130873062-130873277	24.3	0.0139	0.1651	ENSMUSG00000030173	KILLER CELL LECTIN-LIKE RECEPTOR 5 (T-CELL SURFACE GLYCOPROTEIN LY-49E) (LY49-E ANTIGEN).
65	6.17642460-17642693	42.4	0.0527	0.0263	ENSMUSG00000029534	SUPPRESSION OF TUMORIGENICITY 7
66	6.18782659-18784598	30.3	0.0328	0.0633	ENSMUSG00000029517	NA
67	6.23207766-23211246	152.8	0.0743	1.1548	ENSMUSG00000029697	NA
68	6.52593717-52596796	112.1	0.0045	1.3065	ENSMUSG00000029844	HOMEOBOX PROTEIN HOX-A1 (HOX-1.6) (HOMEOTIC PROTEIN ERA-1-993) (EARLY RETINOIC ACID 1) (HOMEOBOXLESS PROTEIN ERA-1-399).
69	6.52596553-52596796	4.8	0.0038	0.0221	ENSMUSG00000038243	HOMEOBOX PROTEIN HOX-A6 (HOX-1.2) (M5-4).
70	6.52601791-52603517	56.4	0.0038	0.6439	ENSMUSG00000000942	HOMEOBOX PROTEIN HOX-A4 (HOX-1.4) (MH-3).

					ENSMUSG00000014704	HOMEOBOX PROTEIN HOX-A2 (HOX-1.11).
					ENSMUSG00000038243	HOMEOBOX PROTEIN HOX-A6 (HOX-1.2) (M5-4).
					ENSMUSG00000038253	HOMEOBOX PROTEIN HOX-A5 (HOX-1.3) (M2).
					ENSMUSG00000038270	HOMEOBOX PROTEIN HOX-A3 (HOX-1.5) (MO-10).
71	6.52606992-52612049	343.9	0.0055	4.0843	ENSMUSG00000000942	HOMEOBOX PROTEIN HOX-A4 (HOX-1.4) (MH-3).
					ENSMUSG00000038236	HOMEOBOX PROTEIN HOX-A7 (HOX-1.1) (M6-12) (M6).
					ENSMUSG00000038243	HOMEOBOX PROTEIN HOX-A6 (HOX-1.2) (M5-4).
					ENSMUSG00000038253	HOMEOBOX PROTEIN HOX-A5 (HOX-1.3) (M2).
					ENSMUSG00000038270	HOMEOBOX PROTEIN HOX-A3 (HOX-1.5) (MO-10).
72	6.52614853-52615076	8.2	0.0107	0.0008	ENSMUSG00000000942	HOMEOBOX PROTEIN HOX-A4 (HOX-1.4) (MH-3).
					ENSMUSG00000038227	HOMEOBOX PROTEIN HOX-A9 (HOX-1.7).
					ENSMUSG00000038236	HOMEOBOX PROTEIN HOX-A7 (HOX-1.1) (M6-12) (M6).
					ENSMUSG00000038243	HOMEOBOX PROTEIN HOX-A6 (HOX-1.2) (M5-4).
					ENSMUSG00000038253	HOMEOBOX PROTEIN HOX-A5 (HOX-1.3) (M2).
73	6.52615998-52621268	262.3	0.0270	2.9054	ENSMUSG00000000942	HOMEOBOX PROTEIN HOX-A4 (HOX-1.4) (MH-3).
					ENSMUSG00000038227	HOMEOBOX PROTEIN HOX-A9 (HOX-1.7).
					ENSMUSG00000038236	HOMEOBOX PROTEIN HOX-A7 (HOX-1.1) (M6-12) (M6).
					ENSMUSG00000038243	HOMEOBOX PROTEIN HOX-A6 (HOX-1.2) (M5-4).
					ENSMUSG00000038253	HOMEOBOX PROTEIN HOX-A5 (HOX-1.3) (M2).
74	6.61427657-61427934	41.3	0.0212	0.3026	ENSMUSG00000025891	NA
75	6.61465369-61465636	21.4	0.0253	0.0249	ENSMUSG00000025891	NA
76	6.6626082-6628712	129.1	0.0668	0.9397	ENSMUSG00000029755	HOMEOBOX PROTEIN DLX-5.
77	6.68701932-68702134	11.1	0.0130	0.0142	ENSMUSG00000029895	IG KAPPA CHAIN V-II REGION VKAPPA167 PRECURSOR.
78	6.78137598-78139829	86.7	0.1026	0.1016	ENSMUSG00000030026	ALPHA-2 CATENIN (ALPHA-CATENIN RELATED PROTEIN) (ALPHA N-CATENIN).
					ENSMUSG00000037682	NA
79	6.89060972-89067216	180.9	0.1074	1.1894	ENSMUSG00000015053	ENDOTHELIAL TRANSCRIPTION FACTOR GATA-2.
80	6.90977417-90978167	23.5	0.0228	0.0730	ENSMUSG00000034468	VN5 (VOMERONASAL RECEPTOR V1RB3).
81	7.127391482-127405249	396.4	0.0527	4.2814	ENSMUSG00000010476	TRANSCRIPTION FACTOR COE3 (EARLY B-CELL FACTOR 3) (EBF-3) (OLF-1/EBF- LIKE 2) (OE-2) (O/E-2).
82	7.13599839-13600074	10.8	0.0121	0.0196	ENSMUSG00000005602	MYOTUBULARIN-RELATED PROTEIN 2 (FRAGMENT).
					ENSMUSG00000008991	CEA13 PROTEIN (FRAGMENT).
83	7.19240894-19241512	46.1	0.0242	0.3316	ENSMUSG00000003017	CYTOCHROME P450 2A12 (EC 1.14.14.1) (CYPIIA12) (STEROID HORMONES 7- ALPHA-HYDROXYLASE) (TESTOSTERONE 7-ALPHA-HYDROXYLASE).
84	7.38114850-38115524	22.8	0.0185	0.1045	ENSMUSG00000030476	G PROTEIN-COUPLED RECEPTOR.
85	7.39256410-39259340	205.4	0.1246	1.3264	ENSMUSG00000030507	HOMEOBOX PROTEIN DBX.
86	7.46828784-46828999	33.8	0.0329	0.1046	ENSMUSG00000030449	GAMMA-AMINOBUTYRIC-ACID RECEPTOR GAMMA-3 SUBUNIT PRECURSOR (GABA(A) RECEPTOR).
87	7.6640131-6640338	6.8	0.0075	0.0127	ENSMUSG00000000605	CHLORIDE CHANNEL PROTEIN 4 (CLC-4).

					ENSMUSG00000034155	ADULT MALE TESTIS CDNA, RIKEN FULL-LENGTH ENRICHED LIBRARY, CLONE=4930547K11, FULL INSERT SEQUENCE.
88	7.94712714-94712935	14.7	0.0141	0.0481	ENSMUSG00000041946	MITOGEN-ACTIVATED PROTEIN KINASE KINASE 1 INTERACTING PROTEIN 1
89	8.100420881-100421409	13.2	0.0156	0.0159	ENSMUSG00000035880	RNA FOR TYPE IIB INTRACISTERNAL A-PARTICLE (IAP) ELEMENT ENCODING INTEGRASE, CLONE 106 (IAP) (RNA FOR TYPE IIB INTRACISTERNAL A-PARTICLE (IAP) ELEMENT ENCODING INTEGRASE, CLONE 111).
90	8.100424459-100424746	17.3	0.0170	0.0527	ENSMUSG00000035880	RNA FOR TYPE IIB INTRACISTERNAL A-PARTICLE (IAP) ELEMENT ENCODING INTEGRASE, CLONE 106 (IAP) (RNA FOR TYPE IIB INTRACISTERNAL A-PARTICLE (IAP) ELEMENT ENCODING INTEGRASE, CLONE 111).
91	8.104629960-104630191	26.1	0.0252	0.0824	ENSMUSG00000031884	SIMILAR TO CARBOXYLESTERASE 2 (INTESTINE, LIVER).
92	8.104661963-104662473	19.8	0.0196	0.0585	ENSMUSG00000031884	SIMILAR TO CARBOXYLESTERASE 2 (INTESTINE, LIVER).
93	8.128411535-128414389	97.9	0.1151	0.1211	ENSMUSG00000025810	NEUROPILIN-1 PRECURSOR (A5 PROTEIN).
94	8.21535234-21538161	142.4	0.1003	0.7916	ENSMUSG00000031539	ADAPTOR-RELATED PROTEIN COMPLEX AP-3 MU2 SUBUNIT.
95	8.44172592-44173463	80.1	0.0570	0.4398	ENSMUSG00000031646	MOUSE FAT 1 CADHERIN (FRAGMENT).
96	8.44186073-44189712	186.8	0.0559	1.7321	ENSMUSG00000031646	MOUSE FAT 1 CADHERIN (FRAGMENT).
					ENSMUSG00000038952	MELATONIN RECEPTOR TYPE 1A (MEL-1A-R).
97	8.44498250-44499005	41.1	0.0498	0.0373	ENSMUSG00000031640	PLASMA KALLIKREIN PRECURSOR (EC 3.4.21.34) (PLASMA PREKALLIKREIN) (KININOGENIN) (FLETCHER FACTOR).
98	8.91555744-91561784	43.0	0.0559	0.0041	ENSMUSG00000031734	IROQUOIS-CLASS HOMEODOMAIN PROTEIN IRX-3.
99	8.9936844-9938878	74.1	0.0416	0.5093	ENSMUSG00000031497	TUMOR NECROSIS FACTOR LIGAND SUPERFAMILY MEMBER 13B (B CELL-ACTIVATING FACTOR) (BAFF).
					ENSMUSG00000040396	NA
100	9.38502882-38503084	14.2	0.0157	0.0258	ENSMUSG00000040248	PUTATIVE OLFACTORY RECEPTOR (FRAGMENT).
101	9.56065428-56072433	236.8	0.0984	1.9437	ENSMUSG00000032314	SIMILAR TO ELECTRON-TRANSFER-FLAVOPROTEIN, ALPHA POLYPEPTIDE (GLUTARIC ACIDURIA II).
102	9.75504124-75505553	90.4	0.1004	0.1658	ENSMUSG00000034924	HEPATOCYTE NUCLEAR FACTOR 6 (HNF-6) (ONE CUT DOMAIN FAMILY MEMBER 1).
103	9.87409038-87412203	94.3	0.0918	0.2909	ENSMUSG00000032415	NA
104	9.91848564-91852357	139.2	0.0297	1.4004	ENSMUSG00000032368	ZINC FINGER PROTEIN ZIC1 (ZINC FINGER PROTEIN OF THE CEREBELLUM 1).
					ENSMUSG00000036972	ZINC FINGER PROTEIN ZIC4 (ZINC FINGER PROTEIN OF THE CEREBELLUM 4).
105	9.91867752-91870711	155.4	0.0368	1.5305	ENSMUSG00000036972	ZINC FINGER PROTEIN ZIC4 (ZINC FINGER PROTEIN OF THE CEREBELLUM 4).
106	10.103417472-103417775	45.0	0.0474	0.1053	ENSMUSG00000019892	NA
107	10.121627379-121630233	152.2	0.1359	0.5833	ENSMUSG00000034707	NA
108	10.128394146-128398404	67.5	0.0641	0.2238	ENSMUSG00000025399	RETINOL DEHYDROGENASE TYPE 6.
109	10.25395244-25395985	242.2	0.0451	2.4982	ENSMUSG00000019978	BAND 4.1-LIKE PROTEIN 2 (GENERALLY EXPRESSED PROTEIN 4.1) (4.1G).

110	10.35743727-35743941	17.3	0.0199	0.0252	ENSMUSG00000039439	S-ADENOSYLMETHIONINE DECARBOXYLASE PROENZYME 1 (EC 4.1.1.50) (ADOMETDC 1) (SAMDC 1)
111	10.41525848-41527854	99.9	0.1062	0.2267	ENSMUSG00000019821	NA
112	11.11316328-11316706	22.1	0.0240	0.0451	ENSMUSG00000020193	ZONA PELLUCIDA BINDING PROTEIN.
113	11.21593389-21594262	35.5	0.0433	0.0296	ENSMUSG00000020321	MALATE DEHYDROGENASE, CYTOPLASMIC (EC 1.1.1.37).
114	11.5889970-5892641	98.1	0.0538	0.6851	ENSMUSG00000020465	CALMODULIN-DEPENDENT PROTEIN KINASE II BETA M ISOFORM (FRAGMENT).
					ENSMUSG00000020466	CALCIUM/CALMODULIN-DEPENDENT PROTEIN KINASE TYPE II BETA CHAIN (CAM- KINASE II BETA CHAIN) (EC 2.7.1.123) (CAMK-II, BETA SUBUNIT).
115	11.59486736-59487762	47.8	0.0596	0.0273	ENSMUSG00000020455	TRIPARTITE MOTIF PROTEIN TRIM11.
					ENSMUSG00000020496	UNKNOWN PROTEIN (FRAGMENT).
					ENSMUSG00000036952	TRIPARTITE MOTIF PROTEIN TRIM17 (FRAGMENT).
116	11.97047828-97048047	20.9	0.0262	0.0107	ENSMUSG00000000690	HOMEOBOX PROTEIN HOX-B6 (HOX-2.2) (MH-22A).
					ENSMUSG00000038684	HOMEOBOX PROTEIN HOX-B3 (HOX-2.7) (MH-23).
					ENSMUSG00000038692	HOMEOBOX PROTEIN HOX-B4 (HOX-2.6).
					ENSMUSG00000038700	HOMEOBOX PROTEIN HOX-B5 (HOX-2.1) (MU-1) (H24.1).
117	11.97052217-97054059	17.5	0.0101	0.1179	ENSMUSG00000000690	HOMEOBOX PROTEIN HOX-B6 (HOX-2.2) (MH-22A).
					ENSMUSG00000038684	HOMEOBOX PROTEIN HOX-B3 (HOX-2.7) (MH-23).
					ENSMUSG00000038692	HOMEOBOX PROTEIN HOX-B4 (HOX-2.6).
					ENSMUSG00000038700	HOMEOBOX PROTEIN HOX-B5 (HOX-2.1) (MU-1) (H24.1).
118	11.97056320-97058384	127.1	0.0050	1.4812	ENSMUSG00000038684	HOMEOBOX PROTEIN HOX-B3 (HOX-2.7) (MH-23).
					ENSMUSG00000038692	HOMEOBOX PROTEIN HOX-B4 (HOX-2.6).
					ENSMUSG00000038700	HOMEOBOX PROTEIN HOX-B5 (HOX-2.1) (MU-1) (H24.1).
119	12.108799297-108799519	15.6	0.0123	0.0753	ENSMUSG00000003771	IG MU CHAIN C REGION.
120	12.108908897-108909473	10.8	0.0015	0.1165	ENSMUSG00000003771	IG MU CHAIN C REGION.
					ENSMUSG00000037192	IG HEAVY CHAIN V REGION 36-65.
121	12.109727905-109728117	14.7	0.0084	0.0991	ENSMUSG00000002716	IG HEAVY CHAIN V REGION 23 PRECURSOR.
122	12.109728895-109729173	8.1	0.0084	0.0203	ENSMUSG00000002716	IG HEAVY CHAIN V REGION 23 PRECURSOR.
123	12.51155681-51158627	151.5	0.0559	1.3091	ENSMUSG00000001497	PAIRED BOX PROTEIN PAX-9.
124	12.52685216-52685629	36.9	0.0328	0.1427	ENSMUSG00000035431	SOMATOSTATIN RECEPTOR TYPE 1 (SS1R) (SRIF-2).
125	13.115198010-115198363	15.4	0.0199	0.0020	ENSMUSG00000021730	HYPERPOLARIZATION-ACTIVATED, CYCLIC NUCLEOTIDE-GATED K+ 1.
126	13.21140069-21142250	66.4	0.0509	0.3309	ENSMUSG00000008648	HISTONE H3.1 (H3/A) (H3/C) (H3/D) (H3/F) (H3/H) (H3/I) (H3/J) (H3/K) (H3/L).
					ENSMUSG00000016909	HISTONE H2B 291B.
					ENSMUSG00000016977	HISTONE H2B F (H2B 291A).
					ENSMUSG00000036577	HISTONE H4.
127	13.21467693-21470325	131.9	0.0669	0.9722	ENSMUSG00000006179	THYMUS-SPECIFIC SERINE PROTEASE PRECURSOR (EC 3.4.-.-).
128	13.22831204-22833551	115.4	0.0645	0.7960	ENSMUSG00000036376	ACTIVATOR OF BASAL TRANSCRIPTION.

129	13.22943293-22945160	91.8	0.0954	0.2273	ENSMUSG00000018084	HISTONE H3.1 (H3/A) (H3/C) (H3/D) (H3/F) (H3/H) (H3/I) (H3/J) (H3/K) (H3/L).
					ENSMUSG00000018094	HISTONE H2B F (H2B 291A).
					ENSMUSG00000018100	HISTONE H1.3 (H1 VAR.4) (H1D).
					ENSMUSG00000018101	HISTONE H2A.G (H2A/G) (H2A.3).
					ENSMUSG00000036243	HISTONE H4.
					ENSMUSG00000036253	HISTONE H3 (H3.2).
					ENSMUSG00000036326	HISTONE H3 (H3.2).
130	13.23154492-23156486	51.3	0.0608	0.0590	ENSMUSG00000006611	HEREDITARY HAEMOCHROMATOSIS PROTEIN HOMOLOG PRECURSOR.
					ENSMUSG00000016575	HISTONE H2A.G (H2A/G) (H2A.3).
					ENSMUSG00000036132	HISTONE H1.1 (H1 VAR.3) (H1A).
					ENSMUSG00000036149	HISTONE H3 (H3.2).
					ENSMUSG00000036173	HISTONE H3 (H3.2).
					ENSMUSG00000036201	HISTONE H4.
131	13.23160436-23162013	86.1	0.0585	0.4976	ENSMUSG00000006611	HEREDITARY HAEMOCHROMATOSIS PROTEIN HOMOLOG PRECURSOR.
					ENSMUSG00000036132	HISTONE H1.1 (H1 VAR.3) (H1A).
					ENSMUSG00000036149	HISTONE H3 (H3.2).
					ENSMUSG00000036173	HISTONE H3 (H3.2).
132	13.23170125-23173292	77.3	0.0622	0.3591	ENSMUSG00000036132	HISTONE H1.1 (H1 VAR.3) (H1A).
					ENSMUSG00000036173	HISTONE H3 (H3.2).
133	13.27591935-27593045	41.0	0.0119	0.3837	ENSMUSG00000017064	NA
134	13.52751878-52756081	289.4	0.0930	2.6253	ENSMUSG00000021469	HOMEOBOX PROTEIN MSX-2 (HOX-8.1).
135	13.55311592-55317979	254.6	0.0955	2.1841	ENSMUSG00000021506	PITUITARY HOMEOBOX 1 (HOMEOBOX PROTEIN P-OTX) (PITUITARY OTX-RELATED FACTOR) (HINDLIMB EXPRESSED HOMEOBOX PROTEIN BACKFOOT) (PTX1).
136	13.60496444-60496700	32.0	0.0123	0.2722	ENSMUSG00000021433	CTLA-2-BETA PROTEIN PRECURSOR (FRAGMENT).
					ENSMUSG00000033834	TROPHOBLAST-SPECIFIC PROTEIN PRECURSOR.
137	13.62386305-62396461	335.9	0.0795	3.3092	ENSMUSG00000021466	PATCHED PROTEIN HOMOLOG 1 (PTC1) (PTC).
138	13.69029605-69034427	94.9	0.0219	0.9398	ENSMUSG00000021602	IROQUOIS-CLASS HOMEODOMAIN PROTEIN IRX-1 (FRAGMENT).
139	13.70335945-70338999	91.1	0.0359	0.7664	ENSMUSG00000021604	IROQUOIS RELATED HOMEOBOX 4 (DROSOPHILA).
140	13.75280509-75289493	125.8	0.0364	1.1782	ENSMUSG00000035892	COUP TRANSCRIPTION FACTOR 1 (COUP-TF1) (COUP-TF I).
141	13.75295999-75296808	24.2	0.0208	0.0995	ENSMUSG00000035892	COUP TRANSCRIPTION FACTOR 1 (COUP-TF1) (COUP-TF I).
142	13.88157413-88157619	32.0	0.0397	0.0199	ENSMUSG00000021619	GENE.
143	13.92203897-92209278	320.1	0.0860	3.0592	ENSMUSG00000021685	HOMEOBOX PROTEIN ORTHOPEDIA.
144	13.97261000-97264806	334.5	0.0858	3.2335	ENSMUSG00000021647	COCAINE- AND AMPHETAMINE-REGULATED TRANSCRIPT PROTEIN PRECURSOR
145	14.108328953-108329184	33.1	0.0361	0.0665	ENSMUSG00000032891	GLYPICAN-6 PRECURSOR.
146	14.108529702-108529910	24.0	0.0217	0.0894	ENSMUSG00000032891	GLYPICAN-6 PRECURSOR.
147	14.110165101-110167920	88.0	0.0852	0.2761	ENSMUSG00000003953	NA
					ENSMUSG00000022133	NA
					ENSMUSG00000022136	DNAJ (HSP40) HOMOLOG, SUBFAMILY C, MEMBER

						3
148	14.110692385-110692613	19.5	0.0233	0.0201	ENSMUSG00000042021	HEPARAN SULFATE 6-O-SULFOTRANSFERASE 3.
149	14.113365106-113368030	137.8	0.1231	0.5262	ENSMUSG00000025544	TRANSMEMBRANE 9 SUPERFAMILY PROTEIN MEMBER 2 PRECURSOR.
150	14.115910243-115910519	29.5	0.0184	0.1856	ENSMUSG00000025551	FIBROBLAST GROWTH FACTOR-14 (FGF-14) (FIBROBLAST GROWTH FACTOR HOMOLOGOUS FACTOR 4) (FHF-4).
151	14.17101881-17113285	430.8	0.1062	4.2038	ENSMUSG00000021778	NA
152	14.23611440-23618208	147.0	0.0444	1.3596	ENSMUSG00000021994	WNT-5A PROTEIN PRECURSOR.
153	14.44583671-44584386	17.2	0.0033	0.1771	ENSMUSG00000022164	T-CELL RECEPTOR ALPHA CHAIN V REGION 2B4 PRECURSOR.
154	14.49205978-49210930	77.4	0.0590	0.3889	ENSMUSG00000021974	GLIA-ACTIVATING FACTOR PRECURSOR (GAF) (FIBROBLAST GROWTH FACTOR-9) (FGF-9) (HBGF-9).
155	14.58366810-58371481	135.6	0.1121	0.6024	ENSMUSG00000022053	TRANSCRIPTION FACTOR COE2 (EARLY B-CELL FACTOR 2) (EBF-2) (OLF-1/EBF- LIKE 3) (OE-3) (O/E-3) (METENCEPHALON-MESENCEPHALON-OLFACTORY TRANSCRIPTION FACTOR 1) (MET-MESENCEPHALON-OLFACTORY TRANSCRIPTION FACTOR 1) (MET-MESENCEPHAL
156	14.70755874-70761803	88.8	0.0998	0.1523	ENSMUSG00000036422	PROTOCADHERIN 8.
157	14.81042676-81043614	30.7	0.0211	0.1758	ENSMUSG00000035043	SIMILAR TO POLY(A)-BINDING PROTEIN, CYTOPLASMIC 4 (INDUCIBLE FORM).
158	14.9636744-9637251	19.1	0.0100	0.1388	ENSMUSG00000021734	INTERLEUKIN-3 RECEPTOR CLASS II ALPHA CHAIN PRECURSOR.
159	14.9637808-9638354	11.9	0.0100	0.0512	ENSMUSG00000021734	INTERLEUKIN-3 RECEPTOR CLASS II ALPHA CHAIN PRECURSOR.
160	14.9643298-9643587	9.7	0.0100	0.0250	ENSMUSG00000021734	INTERLEUKIN-3 RECEPTOR CLASS II ALPHA CHAIN PRECURSOR.
161	15.103911997-103914453	55.3	0.0505	0.2014	ENSMUSG00000001656	HOMEOBOX PROTEIN HOX-C12 (HOX-3.8) (FRAGMENT).
					ENSMUSG00000022484	HOMEOBOX PROTEIN HOX-C10 (HOX-3.6).
					ENSMUSG00000036139	HOMEOBOX PROTEIN HOX-C9 (HOX-3.2).
162	15.103931300-103932309	8.5	0.0100	0.0097	ENSMUSG00000001656	HOMEOBOX PROTEIN HOX-C12 (HOX-3.8) (FRAGMENT).
					ENSMUSG00000001657	HOMEOBOX PROTEIN HOX-C8 (HOX-3.1) (M31).
					ENSMUSG00000022484	HOMEOBOX PROTEIN HOX-C10 (HOX-3.6).
					ENSMUSG00000036139	HOMEOBOX PROTEIN HOX-C9 (HOX-3.2).
163	15.103943516-103943743	12.4	0.0008	0.1416	ENSMUSG00000001657	HOMEOBOX PROTEIN HOX-C8 (HOX-3.1) (M31).
					ENSMUSG00000001661	HOMEOBOX PROTEIN HOX-C6 (HOX-3.3) (HOX-6.1).
					ENSMUSG00000022484	HOMEOBOX PROTEIN HOX-C10 (HOX-3.6).
					ENSMUSG00000022485	HOMEOBOX PROTEIN HOX-C5 (HOX-3.4) (HOX-6.2).
					ENSMUSG00000036139	HOMEOBOX PROTEIN HOX-C9 (HOX-3.2).
164	15.103949349-103949573	1.7	0.0008	0.0133	ENSMUSG00000001657	HOMEOBOX PROTEIN HOX-C8 (HOX-3.1) (M31).
					ENSMUSG00000001661	HOMEOBOX PROTEIN HOX-C6 (HOX-3.3) (HOX-6.1).
					ENSMUSG00000022485	HOMEOBOX PROTEIN HOX-C5 (HOX-3.4) (HOX-6.2).
					ENSMUSG00000036139	HOMEOBOX PROTEIN HOX-C9 (HOX-3.2).
165	15.103953855-103954086	6.7	0.0011	0.0705	ENSMUSG00000001657	HOMEOBOX PROTEIN HOX-C8 (HOX-3.1) (M31).
					ENSMUSG00000001661	HOMEOBOX PROTEIN HOX-C6 (HOX-3.3) (HOX-6.1).
					ENSMUSG00000022485	HOMEOBOX PROTEIN HOX-C5 (HOX-3.4) (HOX-6.2).

					ENSMUSG00000036139	HOMEOBOX PROTEIN HOX-C9 (HOX-3.2).
166	15.103986584-103987987	74.8	0.0214	0.7034	ENSMUSG00000001661	HOMEOBOX PROTEIN HOX-C6 (HOX-3.3) (HOX-6.1).
					ENSMUSG00000022485	HOMEOBOX PROTEIN HOX-C5 (HOX-3.4) (HOX-6.2).
					ENSMUSG00000022486	HOMEOBOX PROTEIN HOX-C4 (HOX-3.5).
167	15.10598728-10600168	83.1	0.0762	0.3001	ENSMUSG00000022246	ANKYCORBIN
					ENSMUSG00000022249	NA
168	15.19053083-19053367	29.6	0.0296	0.0843	ENSMUSG00000022321	CADHERIN-10 (T2-CADHERIN) (FRAGMENT).
169	15.35393215-35397100	127.4	0.0794	0.8032	ENSMUSG00000022330	ODD-SKIPPED-RELATED 2 PROTEIN.
170	15.7632763-7638457	62.8	0.0477	0.3174	ENSMUSG00000022144	GLIAL CELL LINE-DERIVED NEUROTROPHIC FACTOR PRECURSOR.
171	15.97339477-97343786	98.9	0.1043	0.2318	ENSMUSG00000033228	NA
172	15.98911550-98914110	94.3	0.0755	0.4413	ENSMUSG00000022483	COLLAGEN ALPHA 1(II) CHAIN PRECURSOR
173	16.19037149-19037370	11.1	0.0124	0.0202	ENSMUSG00000022695	IG LAMBDA-2 CHAIN V REGION PRECURSOR.
174	16.23744310-23746176	67.3	0.0691	0.1751	ENSMUSG00000022508	B-CELL LYMPHOMA 6 PROTEIN HOMOLOG.
175	16.58693209-58693471	95.9	0.0548	0.6501	ENSMUSG00000035107	NA
176	16.76309614-76310261	30.1	0.0246	0.1359	ENSMUSG00000032932	NA
177	16.89421934-89422163	16.8	0.0205	0.0140	ENSMUSG00000040970	NA
					ENSMUSG00000040984	KERATIN-ASSOCIATED PROTEIN 13.
178	16.89431426-89431787	29.5	0.0175	0.1945	ENSMUSG00000040970	NA
					ENSMUSG00000040984	KERATIN-ASSOCIATED PROTEIN 13.
179	16.10022477-10025712	130.3	0.1310	0.3646	ENSMUSG00000038055	MYLE PROTEIN (DEXAMETHASONE-INDUCED PROTEIN).
180	17.13516524-13516735	16.3	0.0202	0.0108	ENSMUSG00000023886	SECRETED MODULAR CALCIUM-BINDING PROTEIN 2.
181	17.31065239-31070426	260.3	0.0352	2.8059	ENSMUSG00000024042	PROBABLE SERINE/THREONINE PROTEIN KINASE SNF1LK (EC 2.7.1.-) (HRT-20) (MYOCARDIAL SNF1-LIKE KINASE).
182	17.31249825-31252102	107.2	0.1327	0.0712	ENSMUSG00000002070	NA
					ENSMUSG00000002076	NA
183	17.35193993-35194721	37.6	0.0464	0.0263	ENSMUSG00000024432	H-2 CLASS I HISTOCOMPATIBILITY ANTIGEN, TLA(B) ALPHA CHAIN PRECURSOR (MHC THYMUS LEUKEMIA ANTIGEN).
					ENSMUSG00000038311	NA
184	17.54741202-54741531	21.3	0.0221	0.0538	ENSMUSG00000024174	NA
185	17.55481990-55485168	52.9	0.0531	0.1487	ENSMUSG00000013236	PROTEIN-TYROSINE PHOSPHATASE, RECEPTOR-TYPE, S PRECURSOR (EC 3.1.3.48) (PROTEIN-TYROSINE PHOSPHATASE SIGMA) (RPTP-SIGMA) (PROTEIN TYROSINE PHOSPHATASE PTPT9) (PTPASE NU-3).
					ENSMUSG00000024201	SIMILAR TO KIAA0677 GENE PRODUCT.
					ENSMUSG00000024203	NA
186	18.37117509-37119362	91.7	0.0939	0.2407	ENSMUSG00000007705	PROTOCADHERIN ALPHA 4
					ENSMUSG00000007706	CADHERIN-RELATED NEURAL RECEPTER 6 (FRAGMENT).
					ENSMUSG00000007707	CADHERIN-RELATED NEURAL RECEPTER 4 (FRAGMENT).

					ENSMUSG00000041677	PROTOCADHERIN ALPHA 6
					ENSMUSG00000041804	PROTOCADHERIN ALPHA 3.
					ENSMUSG00000041843	PROTOCADHERIN ALPHA 2.
					ENSMUSG00000041884	PROTOCADHERIN ALPHA 1.
187	18.37126631-37128370	130.7	0.0935	0.7131	ENSMUSG00000007705	PROTOCADHERIN ALPHA 4
					ENSMUSG00000007706	CADHERIN-RELATED NEURAL RECEPTER 6 (FRAGMENT).
					ENSMUSG00000007707	CADHERIN-RELATED NEURAL RECEPTER 4 (FRAGMENT).
					ENSMUSG00000041677	PROTOCADHERIN ALPHA 6
					ENSMUSG00000041804	PROTOCADHERIN ALPHA 3.
					ENSMUSG00000041843	PROTOCADHERIN ALPHA 2.
188	18.37139991-37141467	83.8	0.1011	0.0799	ENSMUSG00000007705	PROTOCADHERIN ALPHA 4
					ENSMUSG00000007706	CADHERIN-RELATED NEURAL RECEPTER 6 (FRAGMENT).
					ENSMUSG00000007707	CADHERIN-RELATED NEURAL RECEPTER 4 (FRAGMENT).
					ENSMUSG00000041564	CADHERIN-RELATED NEURAL RECEPTER 8 (FRAGMENT).
					ENSMUSG00000041599	PROTOCADHERIN ALPHA 8.
					ENSMUSG00000041677	PROTOCADHERIN ALPHA 6
189	18.37161117-37163029	95.5	0.0948	0.2790	ENSMUSG00000007440	PROTOCADHERIN ALPHA C2.
					ENSMUSG00000007707	CADHERIN-RELATED NEURAL RECEPTER 4 (FRAGMENT).
					ENSMUSG00000041467	CADHERIN-RELATED NEURAL RECEPTER 7 (FRAGMENT).
					ENSMUSG00000041508	CADHERIN-RELATED NEURAL RECEPTER 3 (FRAGMENT).
					ENSMUSG00000041564	CADHERIN-RELATED NEURAL RECEPTER 8 (FRAGMENT).
					ENSMUSG00000041599	PROTOCADHERIN ALPHA 8.
190	18.37192627-37194006	70.0	0.0644	0.2501	ENSMUSG00000007440	PROTOCADHERIN ALPHA C2.
					ENSMUSG00000041467	CADHERIN-RELATED NEURAL RECEPTER 7 (FRAGMENT).
					ENSMUSG00000041508	CADHERIN-RELATED NEURAL RECEPTER 3 (FRAGMENT).
191	18.37207450-37209321	88.5	0.0418	0.6805	ENSMUSG00000007440	PROTOCADHERIN ALPHA C2.
192	18.37917848-37920059	129.7	0.1062	0.5846	ENSMUSG00000024463	PROTOCADHERIN 13.
					ENSMUSG00000039718	PROTOCADHERIN GAMMA A11.
					ENSMUSG00000039802	PROTOCADHERIN GAMMA B7.
					ENSMUSG00000039839	PROTOCADHERIN GAMMA A10.
					ENSMUSG00000039882	PROTOCADHERIN GAMMA A9.
					ENSMUSG00000039933	PROTOCADHERIN GAMMA B5.
193	18.37929138-37930971	126.2	0.1279	0.3433	ENSMUSG00000024463	PROTOCADHERIN 13.
					ENSMUSG00000039718	PROTOCADHERIN GAMMA A11.
					ENSMUSG00000039802	PROTOCADHERIN GAMMA B7.
					ENSMUSG00000039839	PROTOCADHERIN GAMMA A10.
194	18.63501432-63502121	40.0	0.0378	0.1345	ENSMUSG00000040908	NA

195	18.7178424-7179004	41.2	0.0334	0.1889	ENSMUSG00000024280	NA
196	19.12427634-12428001	26.9	0.0340	0.0112	ENSMUSG00000038666	NA
197	19.39769328-39769573	18.2	0.0081	0.1450	ENSMUSG00000042248	SIMILAR TO CYTOCHROME P450, 2C37.
198	19.43499593-43504124	126.9	0.1334	0.3019	ENSMUSG00000025191	HOMEOBOX PROTEIN NKX-2.3.
199	19.56855416-56860208	299.2	0.0691	2.9629	ENSMUSG00000025081	TUDOR DOMAIN CONTAINING PROTEIN 1.
200	19.61327560-61328116	22.4	0.0279	0.0133	ENSMUSG00000041980	GRANULOCYTE-MACROPHAGE COLONY-STIMULATING FACTOR RECEPTOR ALPHA CHAIN PRECURSOR (GM-CSF-R-ALPHA) (GMR).
201	Un.479515-479818	24.0	0.0091	0.2055	ENSMUSG00000004024	EUKARYOTIC TRANSLATION INITIATION FACTOR 1A (EIF-1A) (EIF-4C) (FRAGMENT).
202	Un.87409427-87409678	17.2	0.0177	0.0444	ENSMUSG00000021888	VOMERONASAL 2, RECEPTOR, 11
203	Un.48665809-48666040	16.4	0.0209	0.0054	ENSMUSG00000024674	NA
204	Un.39514237-39514762	30.1	0.0227	0.1535	ENSMUSG00000033412	T-CELL RECEPTOR ALPHA CHAIN V REGION PHDS58 PRECURSOR.
205	Un.35079087-35079499	28.9	0.0199	0.1641	ENSMUSG00000033431	NA
206	Un.35112573-35112843	20.4	0.0203	0.0591	ENSMUSG00000033431	NA
207	Un.35939364-35939637	33.2	0.0337	0.0900	ENSMUSG00000033431	NA
208	Un.36106885-36107112	21.6	0.0208	0.0693	ENSMUSG00000033431	NA
209	Un.36107737-36108176	19.3	0.0207	0.0415	ENSMUSG00000033431	NA
210	Un.31903807-31904298	25.7	0.0240	0.0892	ENSMUSG00000033563	FERRITIN LIGHT CHAIN 2 (FERRITIN L SUBUNIT 2) (FERRITIN SUBUNIT LG).
211	Un.29259500-29259785	11.4	0.0106	0.0399	ENSMUSG00000033647	60S RIBOSOMAL PROTEIN L30.
212	Un.21587787-21588546	18.8	0.0209	0.0342	ENSMUSG00000033844	UNKNOWN (PROTEIN FOR MGC=6827).
213	Un.107047921-107048175	20.3	0.0237	0.0261	ENSMUSG00000035261	NA
214	Un.106292066-106297686	230.9	0.0356	2.4487	ENSMUSG00000035312	ADULT MALE STOMACH CDNA, RIKEN FULL-LENGTH ENRICHED LIBRARY, CLONE=2210018M05, FULL INSERT SEQUENCE (FRAGMENT).
215	Un.101510780-101511277	28.7	0.0345	0.0290	ENSMUSG00000035482	RING1 AND YY1 BINDING PROTEIN
216	Un.19820425-19821185	18.8	0.0177	0.0636	ENSMUSG00000035684	UNKNOWN (PROTEIN FOR MGC=6827).
217	Un.14128600-14128805	21.5	0.0184	0.0897	ENSMUSG00000035746	NA
218	Un.45133993-45134246	15.1	0.0135	0.0578	ENSMUSG00000042132	HISTONE H2B F (H2B 291A).