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 (α = 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