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