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.